MIKE'S BIOSTATISTICS BOOK

Michael R Dohm
Chaminade University

LibreTexts™

Mike’s Biostatistics Book

Michael R Dohm

Chaminade University

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

TABLE OF CONTENTS

11: Power Analysis

12: One-way Analysis of Variance

13: Assumptions of Parametric Tests

14: ANOVA Designs, Multiple Factors

15: Nonparametric Tests

16: Correlation, Similarity, and Distance

Licensing

Disclaimers

Because this is a biostatistics book, many of the examples and problems come from the medical and biomedical research literature. Nothing in this manuscript, or included as part of any accompanying text, web site, or supplemental material to this manuscript, should be construed as an attempt to offer or render a medical opinion or otherwise engage in the practice of medicine. I may be a doctor, but I'm not that kind of doctor.

Financial and opinion disclaimers

This site is solely supported by me and I don't monetize the site. Responsibility for the content rests with me, not my employer, Chaminade University of Honolulu.

Trademark Notice

Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe or endorse. A partial listing of products mentioned or discussed in this book include the following.

The R statistical programming language is implemented in this book; R is freely available under the GNU General Public License.

Microsoft Access, Excel, Word, Microsoft Office, OneDrive, Windows, Windows XP, Windows Vista, Windows 8, Windows 10, and Windows 11 are registered trademarks of Microsoft Corporation.

Finder, iCloud, iPad, iPhone, macOS, Mac OS X, MacBook, MacBook Pro, OS X, Pages, Quartz, QuickTime, and Safari are registered trademarks of Apple Corporation.

LibreOffice is a registered trademark of The Document Foundation.

Chrome, Chromebook, Google Sheets, Google Docs, Google Drive are registered trademarks of Google LLC.

I have included several cartoons from the wonderful xkcd.com series by Randall Munroe. Copyright clearly belongs to xkcd.

Other products mentioned in this eBook are the trademarks of their registered owners.

No-responsibility disclaimer

Coding examples are provided throughout this eBook. No warranty applies and use of any code examples from this eBook and the "use at your own risk" liability disclaimer applies. The author shall not be liable for any loss of data or other direct or indirect losses that may or may not result from use of code presented herein.

Fair use disclaimer

All efforts to cite and reference scholarly works as needed are included, but there may be some content that may inadvertently refer to works not authorized by the copyright holder. However, *Mike's Biostatistics Book* falls under Section 107 of the Copyright Act because it is intended for educational purposes only.

Copyright

If you wish to use this material, please cite as Dohm, Michael R (2020) Mike's Biostatistics Book, biostatistics.letgen.org. However, as the work is not peer-reviewed I would not recommend the e-book as a sole-source reference.

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

Disclaimers and copyright

Disclaimers

Because this is a biostatistics book, many of the examples and problems come from the medical and biomedical research literature. Nothing in this manuscript, or included as part of any accompanying text, web site, or supplemental material to this manuscript, should be construed as an attempt to offer or render a medical opinion or otherwise engage in the practice of medicine. I may be a doctor, but I'm not that kind of doctor.

Financial and opinion disclaimers

This site is solely supported by me and I don't monetize the site. Responsibility for the content rests with me, not my employer, Chaminade University of Honolulu.

Trademark Notice

Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe or endorse. A partial listing of products mentioned or discussed in this book include the following.

> The R statistical programming language is implemented in this book; R is freely available under the GNU General Public License.
>
> Microsoft Access, Excel, Word, Microsoft Office, OneDrive, Windows, Windows XP, Windows Vista, Windows 8, Windows 10, and Windows 11 are registered trademarks of Microsoft Corporation.
>
> Finder, iCloud, iPad, iPhone, macOS, Mac OS X, MacBook, MacBook Pro, OS X, Pages, Quartz, QuickTime, and Safari are registered trademarks of Apple Corporation.
>
> LibreOffice is a registered trademark of The Document Foundation.
>
> Chrome, Chromebook, Google Sheets, Google Docs, Google Drive are registered trademarks of Google LLC.
>
> I have included several cartoons from the wonderful xkcd.com series by Randall Munroe. Copyright clearly belongs to xkcd.

Other products mentioned in this eBook are the trademarks of their registered owners.

No-responsibility disclaimer

Coding examples are provided throughout this eBook. No warranty applies and use of any code examples from this eBook and the "use at your own risk" liability disclaimer applies. The author shall not be liable for any loss of data or other direct or indirect losses that may or may not result from use of code presented herein.

Fair use disclaimer

All efforts to cite and reference scholarly works as needed are included, but there may be some content that may inadvertently refer to works not authorized by the copyright holder. However, *Mike's Biostatistics Book* falls under Section 107 of the Copyright Act because it is intended for educational purposes only.

Copyright

If you wish to use this material, please cite as Dohm, Michael R (2020) Mike's Biostatistics Book, biostatistics.letgen.org. However, as the work is not peer-reviewed I would not recommend the e-book as a sole-source reference.

Preface

Overview

The following pages, loosely called Mike's Biostatistics Book, contain the extended versions of my lectures for BI311 Biostatistics, a biology course I teach at Chaminade University.

The companion site, Mike's Workbook for Biostatistics, provides **homework**, problems and projects to learn-by-doing biostatistics, and several **R tutorials**.

The Biology department faculty require biology majors to take this course. Class standing of students range from sophomores to first year graduate students.

We use and rely heavily on **R**, the open source "language and environment for statistical computing" (R-project [dot] org), and the **R Commander** package by J Fox (Fox 2005; Fox 2016). R Commander allows students to gain confidence with R commands by use of drop down menus to access functions.

The lecture notes are written from my perspective on what matters in a semester-long, first course in Biostatistics: concepts, context and practical advice along with a generous introduction to general linear models. The focus is on applied statistics with reference to other data science skill sets as needed. Concepts are illustrated by examples and multiple choice questions to "test" reader comprehension. Examples in Mike's Biostatistics Book are generally from real data sets in biological or biomedical research. Therefore, context and practice comes from data sets we will work on throughout the semester. The data sets are presented in the accompanying workbook, body titled ***Mike's Workbook for Biostatistics***.

The material presented in *Mike's Biostatistics Book* provide background and the examples needed to complete the problems presented in the course workbook.

- Chapter 2: Statistical reasoning.
  - Workbook: Homework 1: Assumptions.
- Chapter 3 and 4: Exploring data.
  - Workbook: Homework 2A: Measurement Day results.
  - Workbook: Homework 2B: Descriptive statistics.
- Chapter 5: Experimental design.
- Chapter 6: Probability.
  - Workbook: Homework 3: Distributions & Probability.
- Chapter 7: Risk analysis.
  - Workbook: Homework 4: Risk.
- Chapter 8: Inferential statistics.
  - Workbook: Homework 5: Inference.
- Chapter 9: Qualitative (categorical) analyses.
  - Workbook: Homework 6: Chi-square problems.
- Chapter 10 – 19: Quantitative (continuous) analyses.
  - Chapter 10, 12, 13 – Workbook: Homework 7: t-tests and ANOVA.
  - Workbook: Homework 8: Multiway ANOVA.
  - Workbook: Homework 9: Correlation and simple linear regression.
  - Workbook: Homework 10: Multiple linear regression.
  - Chapter 11: Power analysis.
  - Chapter 15: Nonparametric tests.
  - Chapter 19: Distribution-free methods.
- Chapter 20: Additional topics (partial listing).
  - growth curves, dose response.
  - logistic regression.
  - others.

- Statistical tables.

While the intention is to downplay lists of statistical tests in favor of developing statistical reasoning, many of the kinds of tests one comes across are introduced and discussed in this book. The intent is to introduce these tests as special cases of general (or generalized) models from a data analyst's point of view. Think of "k-means clustering", "independent sample t-test", "ANOVA", "linear regression", and the other tests as vocabulary. We understand biology best when we can talk the talk, and the same holds for learning statistics.

The book does not include **machine learning** — systems that can learn and help make decisions from data — and as of September 2023, includes only a short discourse on clustering and dimensionality reduction of data sets.

About this book (and website)

Equations

Equations in the eBook were created with **LaTeX** — a software system used to prepare and format documents — and saved as PNG images, or embedded in text (QuickLaTeX WordPress plug-in). Pages with many images or equations may be slow to load in your browser: in general, to improve browsing experience reduce the number of open tabs and use of additional apps.

References and citations

Introductory text books often lack in-line citations. The absence of in-line citations improves readability, but at the very real expense of giving credit to the original and to providing the reader the opportunity to verify facts and opinions presented. I have tried a balance: first, I included in-line citations to references. One of many remaining tasks for me to improve the book is to complete linkage of in-page citations to reference lists. I have, however, refrained from an exhaustive, dissertation-like reference listing for each point raised in the book. Second, most citations are to open access articles (or articles with pdfs available by judicious search), with the justification of the reader has access to the original material. However, this approach is a form of **citation bias**. Thus, I refer to reference pages as **References and Suggested Readings**, and don't claim *Mike's Biostatistics Book* as an authoritative voice on the subject (cf. discussion in Greenberg 2009 *Bmj* 339).

A note about me

I'm an Associate Professor of Biology at Chaminade University. My PhD was not in statistics, I trained in evolutionary physiology and quantitative genetics. Quantitative genetics is an applied field that depends heavily on use of mathematics and statistics, particularly linear models. I took courses in applied statistics while at the University of Wisconsin, but I would not call my training in statistics thorough or complete. Much of what I know comes from self-study. My strengths in biostatistics, I believe, are in translation of sometimes dense mathematics to direct use and application. Thus, I have developed a direct style to the material that I hope you will find helpful as you work on the material. It also means we won't spend a lot of time with proofs, not because these are unimportant, but because they can side-track from developing your statistical thinking when it comes to data analysis — but primarily because this is not my strength. References are presented in the book to support the algorithms and mathematical foundations of biostatistics and to back claims I make about applied statistics.

Thus, I don't claim that I have all of the answers, nor am I saying that the mathematical foundations are unimportant, far from it. But we have to start somewhere and I elect to spend our time on the concepts in statistics, the why do we do it this way, as opposed to the mechanics of the mathematics, the how do we do it.

What I have learned about statistics comes from publications from many real statisticians; *Mike's Biostatistics Book*, such as it is, stands on on their work. I apologize in advance to any author whose work has not been given proper credit. Mistakes or mischaracterizations are, of course, mine alone.

Other sources of expertise

Mike's Biostatistics Book of lecture notes is intended to provide students with a foundation in biostatistics: the concepts of assumptions, probability, sampling, description, and modeling that support a researcher's ability to advance knowledge in biology. But you will very much benefit from other opinions, other voices. And, as much as I have adopted the online presence, it is hard to beat a book in hand as a guide. Good statistics books retain their value well passed their publication date. Some of the textbooks I have found useful over the years include the following

Introduction and general statistics

Chaterjee S, Price B (1977). *Regression analysis by example*. Wiley Interscience (5th edition now published in 2006)

Glover T, Mitchell K (2008). *Introduction to biostatistics*, 2nd edition. Waveland Press

Norman GR, Streiner DL (2003). *PDQ Statistics*, 3rd edition. BC Decker

Snedecor GW, Cochran WG (1989). *Statistical methods*, 8th edition. Iowa State University Press

Sokal RR, Rohlf (1981). *Biometry*, 2nd edition. WH Freeman (4th edition published in 2011)

Whitlock MC, Schluter D (2008). *The analysis of biological data*. Roberts and Company

Zar J (1999). *Biostatistical analysis*, 4th edition. Prentice Hall (5th edition published in 2011)

Intermediate and advanced books

Abelson RP (1995). *Statistics as principled argumen*t. Taylor & Francis (epub available)

Bulmer MG (1967). *Principles of statistics*. Dover Publications (epub available)

Davidson AC, Hinkley DV (1997). *Bootstrap methods and their application*. Cambridge University Press (epub available)

Edwards AWF (1992). *Likelihood, expanded edition*. Johns Hopkins University Press

Fisher RA (1934). *Statistical methods for research workers*, 5th edition. Oliver and Boyd (The last edition was the 14th)

Härdle W, Simar L (2003). *Applied multivariate statistical analysis*. Springer-Verlag

Lee PM (1989). *Bayesian statistics: An introduction*. Oxford University Press

McCullagh P, Nelder JA (1989). *Generalized linear models*, 2nd edition. Chapman and Hall

Montgomery DC, Peck EA (1992). *Introduction to linear regression analysis*, 2nd edition. John Wiley & Sons (5th edition published in 2013)

Neter J, Wasserman W, Kutner MH (1989). *Applied linear regression models*, 2nd edition. Robert D Irwin (4th edition published in 2003)

Quinn GP, Keough MJ (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

Shao J (2003). *Mathematical statistics*, 2nd ed. Springer Science

Wei WWS (1990). *Time series analysis*. Addison-Wesley (2nd edition published in 2005)

Some of these titles are old!

One of the good things about statistics is that many of the standard statistical applications were developed a long time ago, so “old textbooks” in statistics retain their value. A quick search online will result in many options to purchase one or more of these books for under $10. In addition, most of the books listed above have new editions; where appropriate I have listed the most recent available edition.

What about books on R?

None of the listed books teach R. Between *Mike’s Biostatistics Book* and the companion *Mike’s Workbook for Biostatistics*, several tutorial and lots of worked examples are provided to help you learn how to use R to help statistical work. A quick Google search, e.g., “free online books learn R,” returns thousands of suggested titles. Search “R tutorials,” for millions more. Chances are, if you have a question about how to do some task in R, someone has already solved the task and published code examples for you to borrow (always cite your sources!).

Concluding remarks about these lecture notes

These collected lecture notes will serve as your official textbook – I have tried to make them accurate, informative, and yet balanced between providing too much detail while still providing depth to the presentation. In class, lecture slides will be provided as outline to these more extensive notes. Homework and quizzes support the progress through the notes.

The lecture notes contained in Mike’s Biostatistics Book are very much a work in progress, with some areas more developed than others. If you find areas that make no sense, seem abrupt, or you would like more examples, please do let me know. Your input is

important to improve this textbook; the Discussion Forum on the course website is a good place to do lend your critiques and suggestions.

Like most subjects, one voice is not enough; you will benefit from acquiring a second opinion, either from one or more of the books listed above or from the many online sites on statistics you will find. The good news is that you will find substantial overlap between what I write and other sources you may acquire because the topics we will cover are foundational and my take is mainstream.

However, you will also find some differences in detail. For one example, I have included much more on risk analysis and an epidemiology tilt as compared to many of the the titles listed under the Introduction and General statistics category. For a second example, I give a different perspective on how to work with probability calculations, emphasizing use of natural numbers over frequency calculations. Many examples provided in the book are drawn from data sets created in lab classes you are or will take while you are at Chaminade University: growth curves, dose-response, working with RT-PCR traces, multi-well plate assays and more.

CHAPTER OVERVIEW

1: Getting Started

How to use this book

This eBook is intended to accompany and support students at Chaminade University of Honolulu enrolled in BI 311 Biostatistics, a one-semester introduction to biostatistics. Like all textbooks, the intent is to provide the reader with a guided and interactive presentation about the subject. However, the text should not be taken as the only voice — there are plenty of good textbooks, many of them free, to help you learn statistics. You are encouraged to seek additional help with the material.

Homework and projects

The book is a standalone product, but the purpose of the book is to provide content for my biostatistics course. Homework and projects designed to build confidence with the material are provided in a separate workbook, also available as a free eBook at https://mikeworkbook.letgen.org/. The companion site, Mike's Workbook for Biostatistics, provides homework and projects to learn-by-doing biostatistics.

The websites serve the course

Mike's Biostatistics Book is hosted at biostatistics.letgen.org. Organization at the site is facilitated by the WordPress them wp-gitbook by Tom J. Nowell at https://github.com/tomjn/wp-gitbook. Thank you, Tom!

The course, BI311 Biostatistics, is a CANVAS CMS website is accessible through chaminade.edu. The course website helps me to organize and support the course. BI311 is web-enhanced course, not a blended or hybrid course; that is, online materials are presented to supplement your work in class and do not replace "face-to-face" time.

Students of BI311 enrolled in CANVAS will find lecture slides, help with your computer, help with R statistical programming language and R Commander, a basic statistics GUI that works with R, and an extensive glossary covering statistics and data science terminology on that site. You will submit your work to the CANVAS website. Online quizzes provide rapid feedback and suggestions for further study.

Material on the website is organized to follow the table of contents from Mike's Biostatistics book. For a 16-week semester, the course would be divided into four parts:

Part 1. Chapter 1 — 5 → Exam 1

Part 2. Chapters 6 – 10 → Exam 2

Part 3. Chapters 11 — 15 → Exam 3

Part 4. Chapters 16 — 18 → Exam 4

Book conventions

Chapters are divided into main subject sections. Headings within sections indicate that important concepts follow, concepts you will be expected to understand and demonstrate that understanding on quizzes and exams.

Figures and tables in each chapter start with Figure 1 and Table 1. If a figure from chapter 3 is referred to in chapter 5, then the reference to that figure will be Figure 3.1, with Figure 1 referring to the first figure in Chapter 5.

Equations were written using **LaTeX**; in some cases, equations are presented as images.

Each section also include questions and additional readings, particularly where there are opinions or interpretations of statistical concepts, references are provided. For example, we spend considerable time discussing what the "p value" means. Additional readings either extend the discussion or provide context to the topic.

Like any new subject, a key to your success will be to learn the language. Terms in the text requiring definitions and your attention are in **bold**. Take care with definitions in statistics: while the words are recognizable, their meanings are often distinctly different from common usage.

Parenthetical notes will appear enclosed in a green box (e.g., this note about statistical terms on this page).

Parenthetical notes will be enclosed in green box (e.g., this note about statistical terms on this page).

R code

Throughout the text we will also include relevant R code. R code you type will appear in a sentence as a code block, like `help.start()` , or as a `preformatted` shaded text block, like so

```
help.start()
```

R code you type will be in blue; any output from R will be displayed as red type.

Each section includes worked examples that are presented to illustrate concepts, demonstrate how R can be used, and include interpretations as appropriate.

Questions for you

I have added questions at the end of most book sections to emphasize important concepts or to have you explore more about a subject. Questions range from extensions of concepts introduced in the section to exercises and problems to solve. Questions in the text are generally short answer or require a numerical solution.

Quiz questions are typically multiple choice or True/False. Questions in the book are intended to provide the student with an outline of topics likely to show up on quizzes and exams. On the other hand, the accompanying workbook, *Mike's Workbook for Biostatistics*, provides students with opportunities to conduct more detailed data analysis and to learn about the R language.

1.1: A quick look at R and R Commander

A first look at R

R is a programming language for statistical computing (Venables et al 2009). R is an **interpreted language** — when you run (execute) an R program — the R interpreter intercepts and runs the code immediately through its **command-line interface**, one line at time. Python is another popular interpreted language common in data science. Interpreted languages are in contrast to **compiled languages**, like C++ and Rust, where program code is sent to a compiler to a machine language application.

The following steps user through use of R and R Commander, from installation to writing and running commands. Mike's Workbook for Biostatistics has a ten-part tutorial, A quick look at R and R Commander, which I recommend.

> **Note**
>
> Getting started? By all means rely on Mike's Biostatistics Book and blogs or other online tutorials to point you in the right direction. You'll also find many free and online books that may provide the right voice to get you working with R. However, the best way to learn is to go to the source. The R team has provided extensive documentation, all included as part of your installation of R. In R, run the command `RShowDoc("doc name")` . replace `doc name` with the name of the **R manual** or **R user guide**. For example, the Venables publication is accessed as `RShowDoc("R-intro")` . Similarly, the manual for installation is `RShowDoc("R-admin")` and the manual for R data import/export is `RShowDoc("R-data")` .

Install R

Full installation instructions are available at Install R and for the R Commander package, at Install R Commander. Here, we provide a brief overview of the installation process.

> **Note**
>
> The instructions at Mike's Biostatistics Book assume use of R on a personal computer running updated Microsoft Windows or Apple macOS operating systems. For Linux instructions, e.g., Ubuntu distro, see How to install R on Ubuntu 22.04. For Chromebook users, if you can install a Linux subsystem, then you can also install and run R, although it's not a trivial installation. For instructions to install R see Levi's excellent writeup at levente.littvay.hu/chromebook/. (This works best with Intel-based CPUs — see my initial attempts with an inexpensive Chromebook at Install R).
>
> Another option is to run R in the cloud via service like Google's Colab or CoCalc hosted by SageMath. Both support Jupyter Notebooks, a "web-based interactive computational environment." Neither cloud-based service supports use of R Commander (because R Commander interacts with your local hardware). Colab is the route I'd choose if I don't have access to a local installation of R.

Download a copy of the R installation file appropriate for your computer from one of the Comprehensive R Archive Network (CRAN) mirror site of the r-project.org. For Hawaii, the most convenient mirror site is provided by the folks at RStudio (https://cloud.r-project.org/).

In brief, Windows 11 users download and install the base distribution. MacOS users must first download and install **XQuartz** (https://Xquartz.org), which provides the X Window System needed by R's **GUI** (graphic user interface). Once XQuartz is installed, proceed to install R to your computer. MacOS users — don't forget to drag the R.app to your Applications folder!

Start R

The following is a minimal look at how to use R and R Commander. Please refer to tutorials at Mike's Workbook for Biostatistics (R work, part 1 – 10) to learn use of R and R Commander.

Once R is installed on your computer, start R as you would any program on your computer. Where discussion requires reference to instructions on use of the R programming language, R code (instructions) the user needs to enter at the R prompt are shown in code blocks.

```
Courier New font within a "code block."
```

Until you write your own functions, the general idea is, you enter one set of commands at a time, one line at a time. For example, to create a new variable, `curry.points`, containing points scored by the NBA's Steph Curry during the 2016 playoffs, type the following code at the **R prompt** (displayed as `>`, the "greater than" sign)

```
curry.points <- c(24,6,40,29,26,28,24,19,31,31,11,18,19)
```

and to obtain the mean, or arithmetic average, for `curry.points` at the R prompt type and enter

```
mean(curry.points)
```

Output from R function mean will look like the following

```
[1] 24.42857
```

The R prompt appears in the RGUI as the greater-than typographical symbol ">" at the beginning of a line (Fig. 1.1). The prompt is returned by R to indicate the interpreter is ready to accept the next line of code.

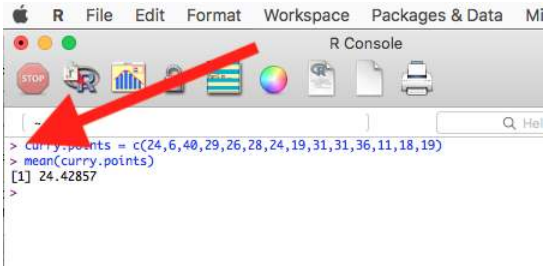


Figure 1.1: The R GUI on a macOS system; red arrow points to the R prompt.

Everything that exists is an object

A brief programmer's note — John M. Chambers, creator of the S programming language and a member of the R-project team, once wrote that sub-header phrase about R and objects. What that means for us: programming objects can be a combination of variables, functions, and data structures. During an R session the user creates and uses objects. The `ls()` function is a useful R command to list objects in memory. If you have been following along with your own installed R app, then how many objects are currently available in your session of R? Answer by submitting `ls()`. Hint: the answer should be one object.

A routine task during analysis is to calculate an estimate then use the result in subsequent work. For example, instead of simply printing the result of `mean(curry.points)`, we can assign the result to an object.

```
myResult <- mean(curry.points)
```

To confirm the new object was created, try `ls()` again. And, of course, there's no particular reason to use the object name, `myResult`, I provided! Like any programming language, creating good object names will make your code easier to understand.

When you submit the above code, R returns the prompt, and the result of the function call is not displayed. View the result by submitting the object's name at the R prompt, in this case, `myResult`. Alternatively, a simple trick is to string commands on the same line by adding ; (semicolon) at the end of the first command. For example,

```
myResult <- mean(curry.points); myResult
```

Write your code as script

While it is possible to submit code one line at a time, a much better approach is to create and manage code in a **script file**. A script file is just a text file with one command per line, but potentially containing many lines of code. Script files help automate R sessions. Once the code is ready, the user submits code to R from the script file.

Note:

Working with scripts eliminates the R prompt, but code is still interpreted one line at a time. The user does not type the prompt in a script file.

Figure 1.2 shows how to create a new script file via the RGUI menu: **File → New script**.

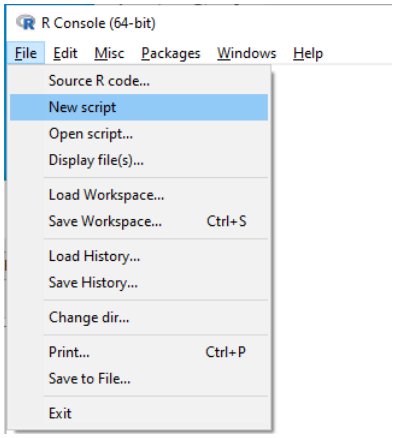


Figure 1.2: Screenshot of drop down menu RGUI, create new script, Windows 10.

The default text editor opens (Fig. 1.3).

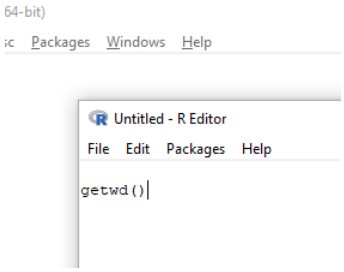


Figure 1.3: Screenshot of a portion of R Script editor, Windows 11. A simple R command is visible.

Submit code by placing cursor at start of the code or, if code consists of multiple lines, selecting all of the code, then hit keyboard keys Ctrl+R (Windows 11) or for macOS, Cmd+Enter.

By default, save R script files for reuse with the file extension .R, e.g., myScript.R. Because the scripts are just text files you can use other editors that may make coding more enjoyable (see RStudio in particular, but there are many alternatives, some free to use. A good alternative is ESS).

Install R Commander package

By now, you have installed the base package of the R statistical programming language. The base package contains all of the components you would need to create and run data analysis and statistics on sets of data. However, you would quickly run into the need to develop functions, to write your own programs to facilitate your work. One of the great things about R is that a large community of programmers have written and contributed their own code; chances are high that someone has already written a function you would need. These functions are submitted in the form of packages. Throughout the semester we will install several R packages to extend R capabilities. R packages discussed in this book are listed at R packages of the Appendix.

Our first package to install is **R Commander**, `Rcmdr` for short. R Commander is a package that adds function to R; it provides a familiar point-and-click interface to R, which allows the user to access functions via a drop-down menu system (Fox 2017). Thus, instead of writing code to run a statistical test, `Rcmdr` provides a simple menu driven approach to help students select and apply the correct statistical test. R Commander also provides access to `Rmarkdown` and a menu approach to rendering reports.

```
install.packages("Rcmdr")
```

In addition, download and install the **plugin**

```
install.packages("RcmdrMisc")
```

See Install R Commander for detailed installation instructions.

Definition:

Plugins are additional software which add function to an existing application.

Start R Commander

After installing Rcmdr, to start R Commander, type `library(Rcmdr)` at the R prompt and enter to load the library

```
library(Rcmdr)
```

On first run of R Commander you may see instructions for installing additional packages needed by R Commander. Accept the defaults and proceed to complete the installation of R Commander. Next time you start R commander the start up will be much faster since the additional packages needed by R Commander will already be present on your computer.

Note that you don't type the R prompt and, indeed, in R Commander **Script window** you won't see the prompt (Fig 1.4). Instead, you enter code in the R Script window, then click "Submit" button (or Win11: Ctrl+R or for macOS: Cmd+Enter), to send the command to the R interpreter. Results are sent to **Output window** (Fig 1.4).

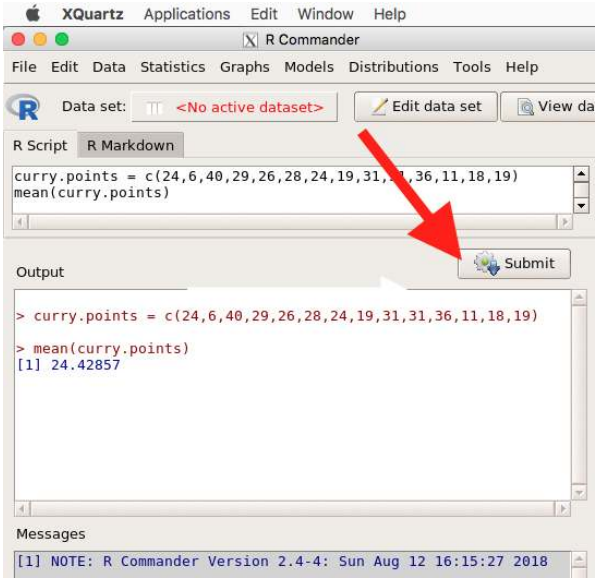


Figure 1.4: The windows of R Commander, macOS. From bottom to top: Messages, Output, Script (tab, Markdown) Rcmdr ver. 2.4-4.

Figure 1.4 shows how the R Commander GUI looks on a macOS computer. The look is similar on Microsoft Windows 11 machines (Fig 1.5).

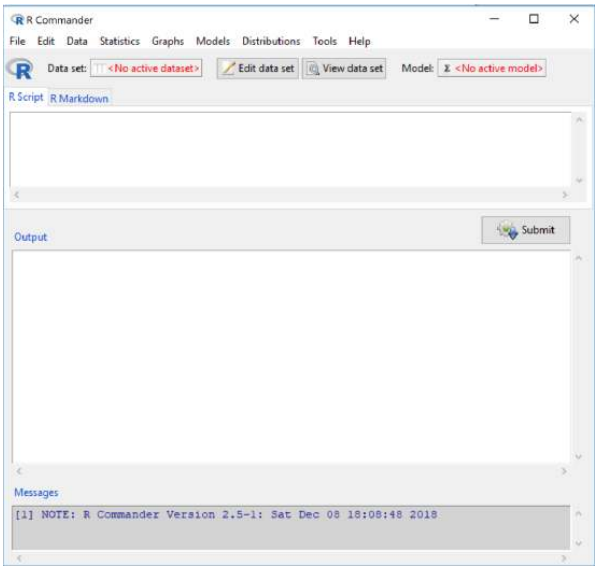


Figure 1.5: The windows of R Commander, Win11. From bottom to top: Messages, Output, Script (tab, R Markdown) Rcmdr ver. 2.5-1.

We use R Commander because it gives us access to code from drop-down menus, which at least initially, helps learn R (Fox 2005, Fox 2016). Later, you'll want to write the code your self, and RStudio provides a nice environment to accomplish your data analysis.

Improve Rcmdr experience. After installing R and Rcdmr, Win11 users should change from MDI to SDI — one big window to separate windows, respectively (see **Do explore settings**, Figure 1.5). macOS users should turn off Apple's app nap (see **Do explore settings**, Figures 1.6 & 1.7), which should improve a Mac user's experience with R Commander and other X Window applications.

Complete R setup by installing LaTeX and pandoc for Markdown

LaTeX is a system for document preparation. pandoc is a document converter system. Markdown is a language used to create formatted writing from simple text code. Once these supporting apps are installed, sophisticated reports can be generated from R sessions, by-passing copy and paste methods one might employ. See Install R Commander for instructions to add these apps.

Note:

If you successfully installed R and are running R Commander, but may be having problems installing pandoc or LaTeX, then this note is for you. While there's advantages to getting pandoc etc working, it is not essential for BI311 work.

Assuming you have Rcmdr and RcmdrMisc installed, and if you have started Rcmdr and have it up and running, then we can skip pandoc and LaTeX installation and use features of your browser to save to pdf.

R Markdown by default will print to a web page (an html document called RcmdrMarkdown.html) and display it in your default browser. To meet requirements of BI311 — you submit pdf files — we can print the html document generated from "Generate Report" in R Commander to a pdf.

- Chrome browser, right click in the web page, from the popup menu select Print, then change destination to Save as pdf.
- Safari browser, right click then select Print page (or if an option, Save page as pdf), then find at lower left find PDF and option to Save as PDF.

R Markdown

Markdown is a syntax for plain text formatting and is really helpful for generating clean html (web) files. R Commander also helps us with our reporting. **R Markdown** is provided as a tab (Fig. 1.4, 1.5). Provided you have also installed `pandoc` on your computer, you can also convert or "render" the work into other formats including pdf and epub. Unsure if your computer has `pandoc` installed? If you are unsure than most likely it is not installed. 😁 `Rcmdr` provides a quick check — go to Tools and if you see Install auxiliary software, then click on it and a link to `pandoc` website to find and download installation file. You can also confirm install of `pandoc` by opening a terminal on your computer (e.g., search " `terminal` " on macOS or " `cmd`

” on Win11), then enter `pandoc –version` at the shell prompt. Figure 1.6 shows version `pandoc` is installed on my Win11 HP laptop.

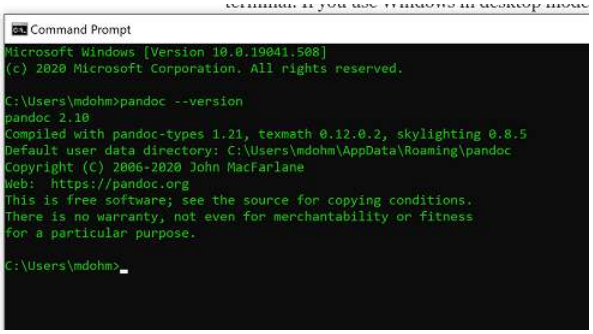


Figure 1.6: Screenshot of terminal window (cmd) on Win11 computer, checking for installed pandoc on a Win10 PC.

Enter your R code in the script window and submit your code, and your results (code, output, graphs) are neatly formatted for you by Markdown. Once the Markdown file is created in R Commander, you can then export to an html file for a a web browser, an MS Word document, or other modes.

Do explore settings!

After installation, R and R Commander are ready to go. However, students are advised that a few settings may need to be changed to improve performance. For example, on Win11 PCs, R Commander recommends changing from the default **MDI (Multiple Document Interface)** to **SDI (Single Document Interface)**. Check the SDI button via Edit menu, select GUI preferences menu. Click save, which will make changes to .RProfile, then exit and restart R. Check to make sure the changes have been made (Fig 1.7).

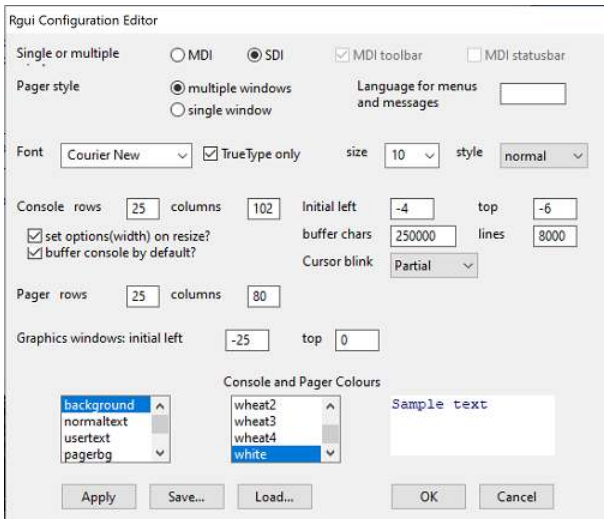


Figure 1.7: Screenshot of GUI preferences settings after changing from default MDI to SDI, Win10.

For macOS users, both R and `Rcmdr` will run better if you turn off Apple’s power saving feature called nap. From `Rcmdr` go to **Tools** and select **Manage Mac OS X app nap for R.app…** (Fig 1.8).

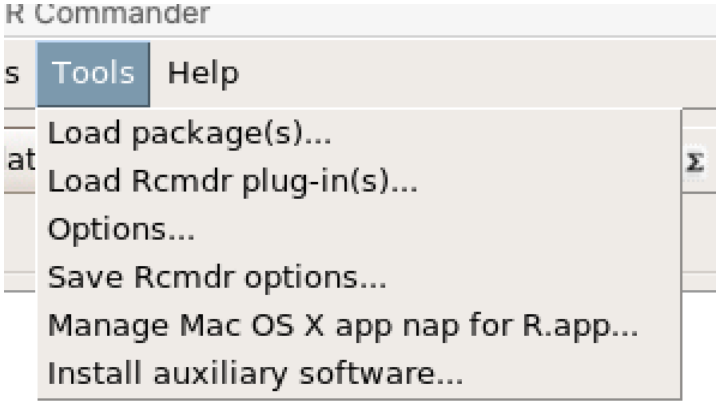


Figure 1.8: Screenshot of Rcmdr Tools popup menu, macOS 10.15.6

A dialog box appears; select off to turn off the app nap (Fig 1.9).

Figure 1.9: Screenshot of Rcmdr Set app nap dialog box, macOS 10.15.6

Exit R Commander

Click on **Rcmdr: File → Exit**, then choose to exit from just R Commander, or both R Commander and R.

If you exit just R Commander or both R and R Commander, you'll receive a pop-up request to confirm you want to quit R Commander (click yes), and a second prompt asking if you want to save your script. In general, select yes and then you'll be able to take up where you left off. Similarly, if asked to save your workspace, choose no. If you save your workspace, this creates an **.RProfile** text file with settings for how R and R Commander will behave the next time you start R. The file will be saved to your current working folder, which R will use the next time it starts. At least while you are getting started, you should avoid creating these .RProfile files.

As long as the current session of R is active, then the library for `Rcmdr`, as well as any other library loaded during the R session, is in memory. To start R Commander again while R is running, at the R prompt, type and submit

```
Commander()
```

Questions

1. Biostatistics students should work through my ten R lessons, called A Quick Look at R and R Commander, available in Mike's Workbook for Biostatistics.
2. Students should also search Internet for R tutorials and R Commander tutorials. Find recent tutorials and work through several of them. We get better when we practice.

1.2: Chapter 1 References and Suggested Readings

Fox, J (2005) The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software* 14(9) https://doi.org/10.18637/jss.v014.i09.

Fox, J (2016). *Using the R Commander: A Point-and-click Interface for R*. Chapman and Hall/CRC.

Laurillard, D (2014). *Five myths about MOOCs. The Times Higher Education https://www.timeshighereducation.com/comment/opinion/five-myths-about-moocs/2010480.article*.

Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL https://posit.co/.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Note:

To get the citation information for R Studio, run the command `RStudio.Version()` in your version of R Studio. Similarly, to cite R, run the command `citation()` in R.

Venables, W. N., Smith, D. M., & R Development Core Team. (2009). An introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics Version 4.3.3 (2024-02-29) https://cran.r-project.org/doc/manuals/R-intro.pdf.

CHAPTER OVERVIEW

2: Introduction

Chapter 1: Getting Started presented a brief introduction to **statistical thinking**, a systematic approach to how we describe and ask questions about the world from data, and a justification for why undergraduate biology students should learn biostatistics. In my day, most of us took statistics as part of our graduate training. The curriculum for science students has accelerated now — it is now assumed that as part of undergraduate career students gain experience working with data and developing quantitative reasoning skills. Biostatistics courses are designed to help you achieve this understanding.

First up, let's sell Why biostatistics?

2.1: Why (Bio)Statistics?

What are the basic elements of biostatistics?

What skills are you going to get from all of this? You will get different opinions on the elements of an essential first course in statistics or biostatistics. Certainly the basics are a foundation in probability and a breadth of classical elementary statistical procedures, which will include descriptive statistics, analysis of variance and linear regression, and an introduction to multivariate analysis. In preparation for your course in epidemiology you will also be introduced to risk analysis and survival analysis. However, the primary return for your time, I hope, will be a deeper appreciation for how to think about problems in biology from experimental design and data analysis perspectives. Practical skills you will learn include how to process and clean data for analysis, data visualization, and a foundation in parametric and nonparametric statistical methods.

Why do we require you to take (Bio)statistics as part of your major?

At Chaminade University we require all biology students to take biostatistics, and we do so with an emphasis on use of data analysis skill development. This requirement aligns our program to national expectations of biology undergraduate education (e.g., AAAS, NAS, NIH, NSF). As stated in *Bio210: Transforming Undergraduate Education for Future Research Biologists*,

> *"Biology majors should be adept at using computers to acquire and process data, carry out statistical characterization of the data and perform statistical tests, and graphically display data in a variety of representations (p. 15)."*

Learning biostatistics from a course like BI-311 — which relies heavily on use of the R programming language and data sets — helps the biology student develop these skills.

In the next pages I will outline a history of statistics (Chapter 2.3), but here I wish to make the point that biostatistics is now considered to be a core skill set for biologists. Biostatistics as a discipline came into its own in the 1930's, but extensive reliance on statistics in research really dates to more recent times because of the ubiquity of personal computers (Salsburg 2002). Modern biological and biomedical research requires computational and quantitative methods to collect, process, analyze, and interpret large data sets. And yet, even a casual survey of required courses in the year 2014 for entry into graduate programs in biology will reveal that biostatistics is not expected of candidates; so what gives?

The first point is that programs list only minimum requirements. The second point is that many programs (genomics, ecology, etc.,) will expect the graduate student to take a year or more of statistics. The need is so crucial that at Harvard Medical School, all biology graduate students are expected to take a crash-course in computing and statistics to work with data (Stefan et al 2015).

Moreover, while graduate programs are not listing statistics as a requirement, many biology undergraduate curricula now require a course in biostatistics to reflect the increasingly data-driven state of modern biology — where the jobs are!

I'll make you a bet — or at least, I'll make this part of your required homework (see BI311 Workbook)! Even a causal search of a research journal article in a biology discipline of your choosing will prove that there is no doing biology research today without an understanding of statistics.

But, you may be thinking, I'm pre-med and plan to apply to medical school …

Even a cursory look at the literature will result in finding many authors strongly calling for this kind of preparation for a successful career in medicine (e.g., Brieger and Hardin 2012). It's obvious, but needs stating — you're applying to medical school to become a doctor — you'll spend the majority of your adult life as a doctor. Statistical thinking is crucial to answering the daily question: "My patient tested positive for biomarker X, what's the chance that the patient has disease Y?" If you answer is, the patient has the disease, then you definitely need this course! Hint: there are four possible outcomes of a test, see Chapter 7.3 – Conditional Probability and Evidence-Based Medicine.

Need more convincing? Take a look at the targets of questions intended to evaluate Skill 4 of the Scientific Inquiry and Reasoning Skills standard of the revised MCAT^{2015} Exam (p. 107, What's on the *MCAT*^{2015} exam?).

- Using, analyzing, and interpreting data in figures, graphs, and tables.
- Evaluating whether representations make sense for particular scientific observations and data.
- Using measures of central tendency (mean, median, and mode) and measures of dispersion (range, inter-quartile range, and standard deviation) to describe data.

- Reasoning about random and systematic error.
- Reasoning about statistical significance and uncertainty (e.g., interpreting statistical significance levels, interpreting a confidence interval).
- Using data to explain relationships between variables of make predictions.
- Using data to answer research questions and draw conclusions.
- Identifying conclusions that are supported by research results.
- Determining the implications of results for real-world situations.

I won't trouble you now with further justifications.

In what disciplines are biostatisticians employed?

One way to begin this discussion is to think about where statisticians work. The job market includes:

Health Science

- Drug design, causes of diseases (many "causes" of cancers).
- Health Professional (nurses, physical therapists).
- Type of care and recovery period (importance of a person's mood on health).
- Exercise regime and recovery from injury.
- Nutrition: vitamins and health, diet and health.

Ecology & Evolution

- Causes of changes in population sizes (conservation biology).
- Effects of pollution on organisms and ecosystems.
- Evolution of traits in populations over time.
- Global environmental changes and changes in population sizes or species diversity.

Genetics & Molecular Biology

- Identifying genes that influence traits (e.g., breast cancer, cystic fibrosis).
- Nature vs. nurture (heredity and environment effects on phenotypes).
- Multiple sequence alignment in comparative genomics.

Agriculture

- Fertilizer effects on plant growth and productivity.
- Compare farming and harvesting methods (e.g., organic vs conventional farming).
- Compare plant hybrids for differences in productivity.

Here's a web site that keeps track of statistics jobs: Jobs in Biostatistics. I would go on to add that experience and competence in statistics would also translate to employment in non-biology fields, e.g., business analytics.

Conclusions

Moving forward, we have much to do — you will be exposed to many specific examples of statistical tests, how to calculate estimators, and how to make inferences from experiments. An important goal of this course is for you to be introduced and develop your ability to design experiments. why should you, as biologists and future health care providers, learn biostatistics?

1. Develop statistical reasoning skills. Most, if not all graduate students will need to take several courses in statistics.
    - Statements about research findings, new and better products, sociological and political issues often depend in large part on some form of statistical analysis.
    - By learning a little about experimental design, sampling, and statistical testing, you will be much closer to being able to participate fully in these debates.
2. Most, if not all graduate students will need to take several courses in statistics.
3. Most, if not all jobs in biology require some training in statistics.

So, there's really no doing biology without at least some knowledge of statistics. You're getting a head start!

Questions

1. Compare the table of contents for *Mike's Biostatistics Book* and our *BI 311 Workbook* against the key terms listed from the MCAT^{2020} Skill 4 expectations. Which chapters do you think cover the key terms in the MCAT expectations?
2. Find and copy definitions for **data processing** and **data cleaning** from
    1. one peer-reviewed, primary source* (e.g., search Google Scholar).
    2. one peer-reviewed, secondary source (e.g., search Google Scholar).
    3. Wikipedia. From these three sources you collected, write your own definitions for data processing and data cleaning.

    (* Not sure what is meant by "sources in science?" Search the phrase in Google 😉)
3. In what field or discipline do you see yourself studying or working by the year 2030? What are the data and analytical skills needed for this field? Cite your source (blogs are fine for this).

2.2: Why do we use R software?

Introduction

Why do we use R Software? Or put another way: Dr D, Why are you *making* me use **R**?

Truth? You can probably use just about any acceptable statistical application to get the work done and achieve the learning objectives we have for beginning biostatistics. However, we will use the **R statistical language** as our primary statistical software in this course. Part of the justification is that all statistical software applications come with a **learning curve**, so you'd start at zero regardless of which application I used for the course. In selecting software for statistics I have several criteria. The software should be:

- free software
- open source
- widely accessible and compatible with ~~all~~ most personal computers
- well-respected and widely used by professionals
- well-supported for the purposes of data analysis and data processing
- really good for making graphics, from the basics to advanced
- capable to handle diverse kinds of statistical tests
- if not exactly easy, the software should have a reasonable learning curve

R meets all of these criteria. **R history** began back in 1993 and has always been available as free software under the terms of the Free Software Foundation's GNU General Public License in source code form. R compiles and runs on a wide variety of UNIX platforms and similar systems, including GNU/LINUX, FreeBSD, and various Linux distros like the popular Ubuntu®, in addition to their more famous Microsoft Windows® and Apple macOS® distributions. To facilitate access to the software, numerous **mirror sites** are available from sites around the world, with cloud.r-project.org supported by RStudio perhaps the most widely used. From December 2021 to December 2022, more than 6 million downloads of base R were made from the RStudio **CRAN mirror** site (CRAN stands for Comprehensive R Archive Network; a mirror refers to a website or server that holds a copy of files from another website/server to make the files available from more than one place).

Note:

One hundred and four mirror sites as of March 2023, 105 different locations (including R CRAN at r-project.org), from which to download R and related packages. Thus, it's not a simple task to count total downloads of R. RStudio has given access to their `changelog` file, which allow one to track numbers of downloads for any package from their mirror site — https://cloud.r-project.org/. Here's the code and recent counts for downloads of R itself (about 400K over a four week period).

```
install.packages("cranlogs")
library(cranlogs)
# How many downloads of base R first four weeks of Fall semester?
 out <- cran_downloads("R", from = "2023-08-21", to = "2023-09-08")
sum(out$count)
```

R output

```
[1] 398524
```

R is straightforward to use once you learn how to work with the language, but has a steep learning curve; after all, it's a programming language. The GUI **R Commander** helps in this process, and eventually, your use of code will become second nature. After the initial growing pains are behind you, **RStudio** likely will be a better solution over R Commander. However, while we need statistical software to do statistics, students in my BI311 course must keep in mind that learning objectives for most biostatistics course are about the concepts and interpretation of statistics, not just use of the software. In other words, learning how to use R is not the focus of BI311 nor will you likely achieve R programming competency by the end of the semester. I certainly

encourage students to strive for competency and I give frequent bonus opportunities to demonstrate coding skills during the semester.

Thus you might ask if the purpose of the course isn't to learn R, why work with R instead of a more familiar app or software, e.g., **Microsoft Excel**® (hereafter simply referred to as Excel), or **Google Sheets**, or even my favorite open-source office alternative, **LibreOffice Calc**? Or, perhaps even just one of the many online calculators, if the course learning objective is to "just" learn about statistics?

First, I believe that real data derived from real biology or biomedical problems are essential elements to a first course in biostatistics. That's not a particularly unique perspective, although I don't have survey results of other statistics instructors to back up the claim. Real problems involve observations on multiple subjects, many variables — large data sets; this alone precludes use of hand calculations and calculators. As a corollary, we will not spend a great deal of time learning the in's and out's of the algorithms that form particular statistical tests. Now, do understand that there is a tremendous benefit to understanding statistics by working through the equations, by looking at the algorithms, and there's no escaping the need for understanding that probability provides the foundation of **statistics inference** (Chapter 8). Thus, for most of us, the statistical software available to us provides an appropriate framework for applying correct statistical tests to our projects. Therefore, the decision is about which statistical package we should use.

Second, R is perhaps *the* choice in academia for statistical software. A PUBMED search found more than 1500 citations of R. Visit Robert A. Muenchen's web page (The popularity of data analysis software, r4stats.com) to see updated statistics on statistical software use. Those of you continuing on to graduate school or to professional schools will find that many of your statistically literate colleagues use R and not one of the commercial programs. While there are many excellent commercial packages (Table 2.2.1), and in some cases you can make spreadsheet programs do statistics (typically add-ins are required), all statistical software come with steep learning curves. Thus, part of my selling point to you is that learning to use R is at the cutting-edge in your field and, given that all of the software you could use can have have their challenges, it is best to work with something that will be around and is in wide use, without the burden of a financial investment.

Table 2.2.1: Comparison of Commercial Statistical Software Programs

Software	Student license?	Limited or full function version	macOS	Windows 11	Fee*	Academic license type
GraphPad Prism	Subscription, $142 per year	Full	Yes	Yes	$202	annual subscription
JMP	Yes, but with purchase of selected textbook	Limited	Yes	Yes	$100	monthly subscription
Minitab	Subscription, $54.99 per year	Full	Yes	Yes	$1610	annual subscription
IBM SPSS	Rental, $76 per year	Full	Yes	Yes	$260	annual rental
SigmaSTAT	No	NA	No	Yes	$299	perpetual
MySTAT	Yes, free	Limited	No	Yes		NA
SYSTAT	No	NA	No	Yes	$739	perpetual
Stata	Subscription, $94 per year	Full	Yes	Yes	$325	annual subscription

last updated November 2022

see Wikipedia for list of additional software

Third, what about online sites like **plot.ly** where, for free, you can plot and, in some cases, calculate statistics? What about the web application at Brightstat, which claims to provide an SPSS-like experience online (Stricker 2008)? While it is true that there are many wonderful websites that can perform many of the statistical tests we will use this semester, these sites are not suitable for more than occasional use.

How to get started with R

The R statistical language, accompanied by additional packages to extend its capabilities beyond basic math and statistical functions, provided a complete statistical environment. R is best viewed as a programming language for statistics (**data analysis**), and **data processing**. Power users of R learn how to write scripts that do t-tests, ANOVA, regression, etc. The scripts are just lines of code that R understands and it provides the user tremendous control over analysis and inference of data sets. Because of this flexibility and power, however, R can be intimidating at first. So, we'll start slowly with scripts, introducing just what we need to get started and build from there. We'll be addressing R issues in more depth over the next several weeks, but for the first week, our goal(s) should be to make sure each of you knows how to start/exit R, how to create and utilize a **working directory**, and how to use R as a calculator. You obtain your copy of R from the R Project for Statistical Computing, available at https://www.r-project.org. Instructions to install R are provided in Install R. A ten-part tutorial to get started using R is provided in Mike's Workbook for Biostatistics.

Note:

A working directory or working folder is something you create on your computer to contain the files and sub-directories of a project. It sets the default location for files you may need to have R read. For example, all of your work for a course (data files, script files, Markdown files), may be stored in a folder called BI311 on your Desktop. For example, on a macOS, the path to the working folder would be

```
/Users/username/Desktop/BI311
```

Why R Commander?

We utilize an R package that provides a menu-driven context to much of the typical statistics one needs to do biostatistics. The package is called R Commander (**Rcmdr**), which provides a graphic user interface or GUI. Rcmdr therefore significantly eases the learning curve for doing statistics with R. We use a package called R Commander, which provides **drop down menus** for most of the typical kinds of analyses. Rcmdr is in use in many courses across the world (more than 20K downloads in September 2023), and among the other GUI available for R, Rcmdr is among the best supported GUI available for R. R Commander function is extended by plug-ins; as of August 2023, there were 36 plugins that extend Rcmdr's capabilities. Instructions to install R Commander are provided in **Install R Commander**.

Note:

Other options to improve use of R include use of RStudio®, which is an integrated development environment, or IDE. RStudio is really nice to use, and happily, you can run R Commander within RStudio. I am also increasingly using shiny apps within the course to help with concept presentation; in the future, I plan to provide a complete shiny app which would allow BI311 students to work interactively with the statistics presented in this text, something like the radiant-rstats project. However, for use in our course, R Commander provides a familiar look as students develop knowledge in the course: simply point and click to access the statistical functions.

Wait! Why don't we use Microsoft Excel? My instructor in {insert course here} used Excel…

A very reasonable question for you to ask — why don't we use Excel or Google Sheets for statistics? Moreover, it is highly likely that you have gained at least some introduction to descriptive statistics and graphing with spreadsheets in former courses — shouldn't we learn statistics within a framework you are already familiar?

After all, "Can't Microsoft Excel do statistics?" Mostly the answer is, no, not really (Fig. 2.2.1).

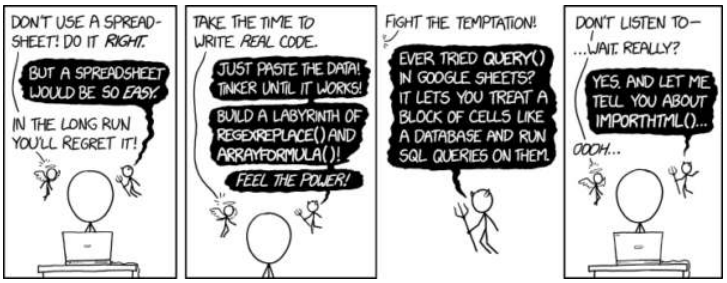


Figure 2.2.1: "Spreadsheets," **xkcd.com** no. 2180

MS Excel, Google Sheets, Apple Numbers, and for that matter, Calc, the spreadsheet application in my favorite office app LibreOffice (LibreOffice is a free, open-source alternative to Microsoft Office), can be used to calculate many descriptive statistics. With some effort, these applications can be extended by use of either Analysis ToolPak or Solver Add-ins to do more complicated statistics like regression and analysis of variance, and curve fitting.

However, use of MS Excel for statistical analysis involves learning a number of commands, syntax, and developing work flows that are neither intuitive nor standard. Some publishers have provided add-ins that are reportedly designed to simplify this process (e.g., MegaStat® by McGraw-Hill or XLStat). None of these options are free and none are in use in any major way by scientists (see The popularity of data analysis software). The free add-ins of Analysis ToolPak and Solver may work for you if you own a Windows PC, but only Solver is included for the Mac versions of Excel. Mac users may download and install StatPlus:MacLE, which is a limited, but free alternative to the Analysis ToolPak add-in; for a complete package a Pro version is available (licenses started at $89, web site: www.analystsoft.com/en/products/statplusmacle/).

An additional caution: you should be aware that there have been reports over the years that algorithms selected by Microsoft for Excel have not always been to industry standards (e.g., McCullogh and Wilson 2005). In short, the fit of Excel and other spreadsheet apps for use in statistics is not a simple one. To do the kinds of statistics we will use routinely in class, Excel would need to be modified with add-ins, and the add-ins would be the result of programming by someone. And you would still need to learn how to write the code.

What about graphics? You may like Microsoft Excel's ability to do graphics. Indeed, Excel, Google Sheets, and LibreOffice Calc can be used to generate many typical kinds of statistical plots. But again, in comparison to R, spreadsheet app graphics are limited and require a deal of effort to generate acceptable plots. I think you'll be surprised at how straight-forward R is. Here's an example, first rendered in Microsoft Excel, then in **base R**. And importantly, the kinds of plots Excel does well at are not necessarily the plots suitable for research publication. For example, Excel allows you to make bar charts easily, but cannot do box plots. Box plots are preferred over bar (column) charts for ratio scale data.

Note:

base R refers to the core R programming language along with many functions and graphics routines. We extend capabilities of base R by adding packages, like R Commander. Definition text

Statistics comparisons between R and MS Excel

About that learning curve. Let's compare R and MS Excel for basic functions common in data analysis. Similar conclusions hold for comparisons to Google Sheets and LibreOffice Calcs. Table 2.2.2 lists the observations we can use to conduct comparisons of the applications.

Table 2.2.2. A simple data set of one variable, A, with 24 observations

varA
12
14
20
25
28

varA
29
32
34
35
39
47
47
50
53
54
71
79
87
89
96
105
122
130
132

One of the first steps in data analysis is to produce what are called descriptive statistics. Common **descriptive statistics** are the **mean** and the **sample standard deviation**. Let's compare Excel and R for retrieving these two statistics.

With Excel, to calculate the arithmetic mean of 24 numbers, enter the values into a single column of 24 rows, then enter " `=average(A2:A25)` ", without the quotes, into a new cell of the spreadsheet. " `A2:A25` " refers to where data would be contained in column `A` rows 2 through 25. Typically the first row in a worksheet would contain the name of the variable, e.g., " `A` ." Depending on the significant figures set, the estimate returned by Excel for the mean of `A` is `59.58333333` .

Similarly, to obtain the standard deviation, type `=stdev(A2:A25)` , into a new cell of the spreadsheet. Again, depending on the significant figures set, Excel returns a value of `37.05215674` for the standard deviation of `A` .

In contrast, to obtain the mean and standard deviation for a variable in an R data set, all you would type at the **R prompt** (`>`), or in the **script window**

Note:

Always run your code as a script. Entering code at the R prompt means you are working at the command-line interface, and you work one line at a time. This is not an efficient way to interact with R. Instead, I recommend you always create and work from a script document. For beginners, that's why I recommend R Commander, which includes a script window. Simply type your code in the script window, highlight the code you wish to run, and run by clicking submit button (or Ctrl+R Win11 or Cmd+Enter macOS). When you are ready to move on from R Commander, RStudio is the IDE of choice.

and then submit, is:

```
A <- c(12, 14, 20, 25, 28, 29, 32, 34, 35, 39, 47, 47, 50, 53, 54, 71, 79, 87, 89, 9(
```

where the “c” is a function to **combine** arguments into a vector and saved to the object `A`, followed at the new line by

```
mean(A)
```

Hit enter after entering the command) and R returns

```
[1] 59.58333
```

For the standard deviation, write the R base function `sd()`

```
sd(A)
```

Hit enter after entering the command and R returns

```
[1] 37.05216
```

It’s not much of a difference, but note that to get the mean (arithmetic average) I typed seven characters in R, but 16 characters in Excel; similarly, for the standard deviation I typed in 5 characters in R, but 13 characters in Excel. That’s a savings of 56% and 62%, respectively. Excel tries to help by using AutoComplete to anticipate what you want to enter, but AutoComplete doesn’t always work properly (e.g., see gene name errors generated by use of default Microsoft Excel settings, Ziemann et al 2016).

Note:

I use spreadsheets all of the time for **data entry** and **data management**. Make sure **AutoComplete** and **AutoCorrect** options are turned off and these problems are much less.

In conclusion, R is quicker for descriptive statistics.

Graphics comparison between R and MS Excel

MS Excel is often cited for its graphics capabilities (Camões 2016). We can make the familiar scatter plots, bar charts, and pie charts in Excel. These plots and more are easily obtained in R. I won’t elaborate here about graphics, since we talk at some length about graphics in Chapter 4. But here’s one example in R.

Let’s plot `B` vs `A`. We already provided the data for variable `A`, here’s the data for variable `B`.

```
17, 21, 21, 26, 27, 32, 28, 42, 40, 30, 71, 53, 56, 61, 55, 89, 82, 63, 116, 162, 116
```

Don’t recall how to assign a set of numbers to an object, B, in R? See above and look again at how we assigned the numbers to object `A`.

To get a simple scatter plot (Fig. 2.2.2), I may write at the R prompt.

```
plot(A,B)
```

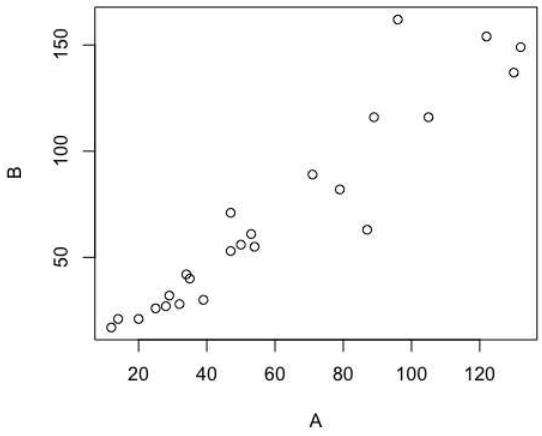


Figure 2.2.2: Basic scatter plot made in R, using `plot(A,B)` .

And here's the comparable default plot (Fig. 2.2.3) from Microsoft Excel, Office 365

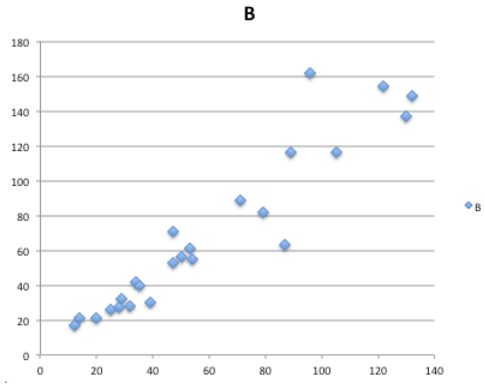


Figure 2.2.3: *Basic scatterplot made in Microsoft Excel.*

Now, both graphs need some work, and to be fair, these are just the defaults. With some effort, you can make an Excel graph look pretty good. But note — the defaults in Excel don't generate axis labels, while R default plot does. Excel adds a useless title and legend; both need to be removed. Excel also adds grid lines where typically one would not include these in a scientific plot.

So, lets count the steps to generate an acceptable scatter plot (Table 2.2.3). I've also added R Commander (`Rcmdr`) steps for comparisons (`Rcmdr` lets you use drop-down menus like Excel or Google Sheets or LibreOffice Calcs).

Table 2.2.3: Steps needed to make a simple scatterplot in R, R Commander, or Microsoft Excel.

Steps	R	Rcmdr	Excel 365
1	write the function	Select Graphs	Highlight columns
2		Select scatterplot	Select from Menu "Insert"
3		Select variables	Select scatterplot
4		Uncheck options	Select type of scatterplot
5			Delete legend
6			Remove grids
7			Insert X-axis label
8			Insert Y-axis label

Conclusion? R is quicker for routine statistical plots like a scatter plot. And I didn't even count the steps needed to change MS Excel's dreadful diamond icon points.

That's one step in R, four steps in `Rcmdr`, but eight steps for Microsoft Excel. LibreOffice Calc is a little better at four steps, but like MS Excel, you'd need to change several components to the graph (Fig. 2.2.4).

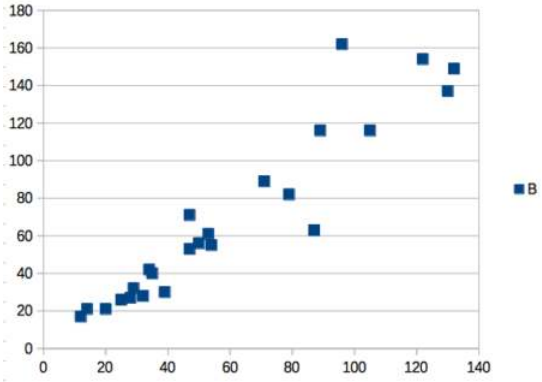


Figure 2.2.4: *Basic scatterplot made in LibreOffice Calc.*

Note:

In R vernacular, these data markers are referred to as **pch**, or **point characters**: `pch = 23` returns a blue diamond character; for a blue square like Figure 2.4, add to the `plot()` command as

```
plot(A,B, pch = 22)
```

So, you're telling me I don't need a spreadsheet application?

No, not at all. We use spreadsheets, and more generally, databases, to store data. Spreadsheets apps are designed to make data entry and data management approachable and efficient. They remain an important tool for researchers (Browman and Woo 2017).

R is not that great of a spreadsheet; packages are available to seamlessly tie your spreadsheet and database data to R via **ODBC**. We will routinely enter and manipulate data in MS Excel, then import the data into R for analysis.

Spreadsheet apps like MS Excel and Google Sheets (see also LibreOffice Calc) are great at being a spreadsheet program; R is great at being a statistical software program. You should take advantage of what the tools do best.

Still not convinced?

R is in use all over the world by students and professionals alike, and if one is going to spend the time to learn how to use a statistics software program, you should learn a standard program, like R.

And it's not just me. Read about R in this 2009 New York Times piece, "Data analysts captivated by R's power." Look who purchased (April 2015) Revolution Analytics, a major player in the development of the R programming language.

Note:

The answer was Microsoft. For several years Microsoft supported R development via Microsoft Machine Learning Server & Microsoft R Open. However, as of July 2023, this service is no longer available. See Microsoft R application network retirement.

Why install R on your computer?

Convenience. Control. Offline.

At the Biology department of Chaminade University, we have installed and maintain R, Rcmdr, and RStudio along with all required packages on our Macbook Pro® Lab computers for your use during class and during optional, proctored biostatistics work sessions. Since 2018, R is increasingly available "in the cloud" (e.g., RStudio Cloud), which would mean you could run R in your browser and avoid installation on your computer. You can run significant analysis with R in the cloud via the free Google Colaboratory and CoCalc platforms that are now available: I encourage you to look into these platforms. Unfortunately, these services are not quite ready for the classroom. For example, RStudio in the Cloud is free to use on a limited basis, but quickly

requires a significant subscription cost with increasing use. Google Colab and GoCalcs require use of Jupyter notebooks, which add yet another layer to the learning curve without focusing on learning statistics. Second, although access to their servers is easy, running simultaneous connections via Chaminade's single public IP address is likely to lead to problems for us. Third, I want you to use R Commander (Rcmdr) to assist in the learning curve — `Rcmdr` cannot be run in the Cloud (i.e., RStudio in the Cloud, Google Colaboratory, or CoCalc).

Therefore, you are encouraged to install R, Rcmdr, and even RStudio onto your own computers, in part because of the convenience, but also because R is not generally available to students on campus, i.e., only the Biology department's computers have the up-to-date R software installed.

To get started, go to your Canvas website and view the file How to install R on your own computer.

An additional benefit to installing a version of R on your computer, you'll understand more about the software if you take the time to install and if need be, troubleshoot your installation of the software. Moreover, there's a considerable amount of help out there for R. For example, a simple Google search (keywords: tutorial "install R"), returns more than 700K hits, and more than 40K January 2023 alone (add "after:2023-01-01" to Google search box). In fact, there's so much out there that you'll want to sample from several sites and select the voice that works best for you.

Questions

1. Conduct the search on Google for tutorials on installing R; find 10 sites and rank them 1 to 10, with 1 being the site you like best and 10 being the one you like least.
    1. For example, I like https://bookdown.org/ndphillips/YaRrr/, which is an online book for working with R and includes detailed instructions for installing R.
2. What are the three reasons I offered to justify use of R over other candidate statistical applications?
3. R may be installed on the public computers available to you in the lab. Check to see if this is true, and if so, what version of R is installed?
4. What does Rcmdr stand for?
5. In your own words, define and contrast GUI applications from IDE applications
6. Try some R work yourself
    1. In R (or Rcmdr), copy and paste the code above for the `A` variable, then create the `B` variable. What happens when you type the variable name by itself at the R prompt?
    2. Make a plot of `A` and `B`, but this time plot `A` against `B`.
        1. What can you conclude about the axis order in the function?

2.3: A brief history of (bio)statistics

Introduction

Before discussing achievements and landmark moments in biostatistics, let's start with basic definitions.

Bioinformatics is loosely defined as a discipline of biology primarily concerned with work involving large data sets (e.g., databases), but a bioinformatician would not primarily be a statistician necessarily. Rather, a bioinformatician, in addition to having a foundation in statistical and mathematical training, would likely be fluent in at least one programming language and confident in the use and design of databases.

Biostatistics, then, refers to use of statistics in biology. Biostatistics encompasses application of statistical approaches to design, analyze, and interpret biological data collected through observation of use of experimentation. In turn, there are many broad disciplines or fields of specialty that trained biostatisticians may work.

Chance, the likelihood that a particular event will occur.

Data scientist, a very general label, is a person likely to work on "big data." Big data may be loosely and inconsistently identified as access to large detailed and unstructured data sets such as visits and behavior within websites of tens of millions of Internet "hits" to a web site like Amazon® or Google®. The data scientist would then be involved in extracting meaning from volumes of this data in a process called **data mining**. In the context of biology, web sites like ALFRED at Yale University that houses allele frequency information collected on human populations, the 1000 genome project, or any of the databases accessible at National Center for Biotechnology Information would constitute sources of big data for biological researchers.

Epidemiology refers to the statistics of patterns of and risk of disease in populations, particularly of humans and thus, an epidemiologist would also be considered to be a biostatistician. The statistics of epidemiology include all of the materials we will cover in this course, but perhaps if any particular analytical approach characterizes epidemiology, it would be **survival analysis** .

Event, an outcome to which a probability is assigned.

Likelihood, the probable chances of occurrence of an event that has already occurred.

Probability, the chance that an event will occur in the future.

Random. This is a good place to share a warning about vocabulary; statistics, like most of science, uses familiar words, but with refined and sometimes different meanings than our every day usage. Consider our everyday use of "random": "without definite aim, direction, rule, or method – subjects chosen at random" (Merriam-Webster online dictionary). In statistics, however, "random" refers to "an assignment of a numerical value to each possible outcome of an event" (Wikipedia). Thus, in statistics, random dictates a method of determining how likely a subject is to be included: If N represents the size of the population, then **random sampling** implies that each individual had $1/N$ chance of being selected. Thus, if $N = 100$, then each individual has a 1% $(1/100)$ chance of selection. This is quite different from Merriam-Webster's definition, in which no method is assumed. To a statistician, then, "random" as used in everyday conversation would imply **haphazard sampling** or **convenience sampling** from a population.

Statistics may be defined as the science of collecting, organizing, and interpreting data. Statistics is a branch of applied mathematics. Note that the word **statistic** is also used, but refers to a calculated quantity like the mean or standard deviation. A little confusing, but the context in which statistics or statistic is appropriate is usually not a major issue.

Some notes about history

The concepts of chance and probability, so crucial to **statistical reasoning**, were realized rather late in the history of mathematics. While people have been writing about applied and theoretical math for thousands of years, probability as a topic of interest by scholars seems to date only back to the late 17th century, beginning with letters written between Pierre de Fermat (1601-1665) and Blaise Pascal (1623 – 1662) and the substantial work on probability by Pierre-Simon Laplace (1749-1827). Often, research on probability developed under the watchful eyes of rich patrons more interested in gaming than to scientific applications. Work on permutations and combinations, essential for an understanding of probability, trace to India prior to Pascal's work (Raju 2011).

The history of statistics goes back further if you allow for the dual use of the term "statistics", both as a descriptor of the act of collecting data and as a systematic approach to the analysis of data. Prior to the 1700s, statistics was used in the sense of collection of data for use by the governments. It is not until the latter part of the 19th century that we see scholarship on statistical analytical techniques. Many of the statistical approaches we teach and use today were developed in the decades between 1880s and the 1930s.

For example, see the work by Francis Galton, Karl Pearson, R. A. Fisher, Sewell Wright, Jerzy Neyman, and Egon Pearson (Karl Pearson's son).

Since the 1950s, there has been an explosion of developments in statistics, particularly as related to power of computers. These include use of resampling, simulation, and Monte Carlo methods (Harris 2010). **Resampling** — the creation of new new samples based on a set of observed data — in particular is a key innovation in statistics. Its use led to a number of innovative ways to estimate the precision of an estimate (see Chapter 3.4 and Chapter 19). **Monte Carlo methods**, or MCM, which involves resampling from a probability distribution, is used to repeat (simulate) an experiment over and over again (Kroese et al 2014). Computers have so influenced statistics that some now define statistics as "…the study of algorithms for data analysis" (p. 175, Beran 2003). For more on the history of statistics, see Anderson (1992), Fienberg (1992), and Freedman 1999; for excellent, conversational books read Salsburg (2002) and McGrayne (2011). For influential women in early development of statisticians, see Anderson (1992).

Epidemiology

John Snow (1813-1858) is credited by some as the "Father of Epidemiology"(Ramsay 2006). During a London outbreak of cholera in 1953, Snow conducted work to establish cholera mortality with source and quality of drinking water. At the time, the prevailing explanation for cholera was that it was an airborne infection. Snow's map of cholera mortality in the Golden Square district of London in relation to a water pump on Broad Street is shown in Fig. 2.3.1. Snow's theory of contaminated water was not accepted as an explanation for cholera until after his death.

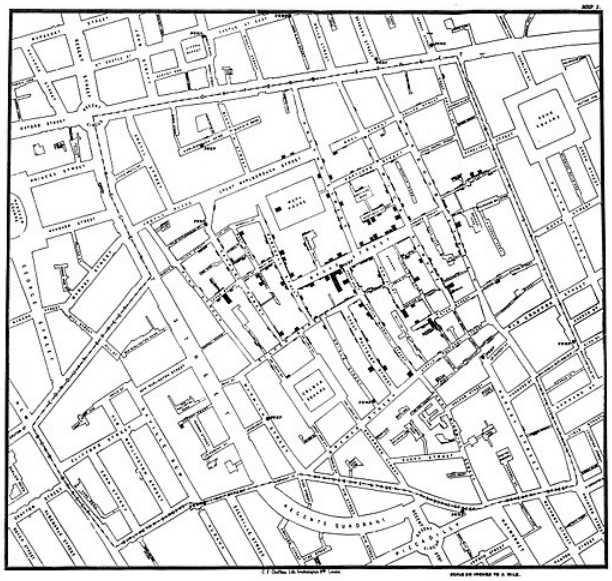


Figure 2.3.1: Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854, drawn and lithographed by Charles Cheffins. Image Public Domain, from Wikipedia

Snow's work and dataset can be viewed and thanks to Paul Lindman and others, the work can be expanded: for example, defining areas around pumps by walking distance (Fig. 2.3.2). The R package is `cholera` . Figure 2.6 shows a plot like Snow's annotated map.cholera.

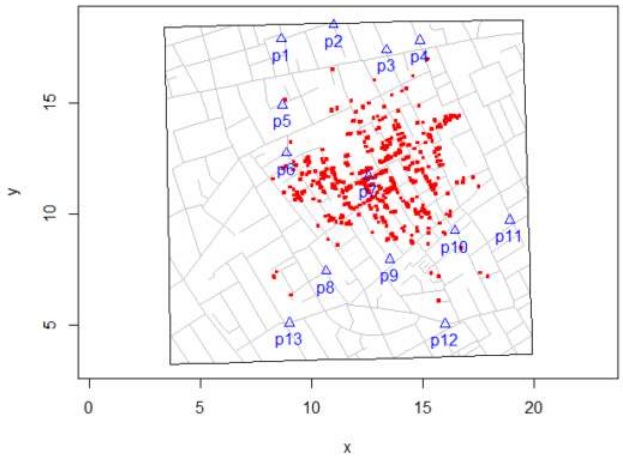


Figure 2.3.2: *Plot of Snow's London using R* `cholera` *package. Triangles marked with p1-p13 represent public water pumps. Red dots represent cholera cases.*

The R code to make the plot was

```
snowMap()
```

Snow's ideas about cholera were not accepted in his time and you should recognize that by itself, a cluster map supports both the airborne and waterborne theories. The `cholera` package contains additional data to help visualize the area, including setting regions by walking distance (Fig. 2.3.3).

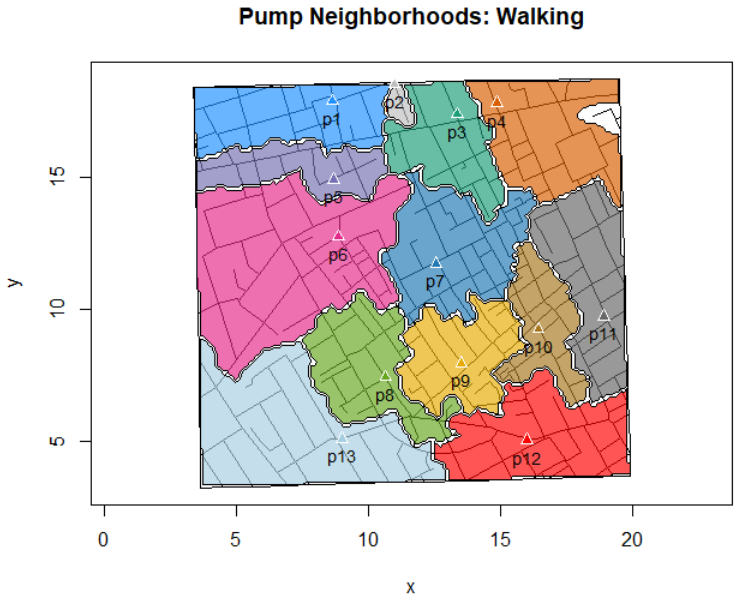


Figure 2.3.3: *Plot of Snow's London with walking areas drawn about the 13 water pumps. Created using R* `cholera` *package.*

R code to make the plot was

```
plot(neighborhoodWalking(case.set = "expected"), "area.polygons")
```

Epidemiology of cancer

This is an undeveloped section in my book. For now, please see Greenwald and Dunn (2009). Key landmarks in the history of epidemiology include

- Tobacco as a carcinogen
- Diet and cancer risk
- Obesity, exercise, and cancer risk
- Hormones and cancer risk
- Cancer risk and occupations: Ramazzini (1713), Pott (1775)

History of founders of statistics and eugenics

Many statistical methods in use today, including regression and analysis of variance methods, can trace their origins to the late 1800's and early 1900's (Kevles 1998). Many of these early statisticians developed statistical methods to further their interests in understanding differences between racial groups of humans. Sir Francis Galton, who developed regression and correlation concepts (the details and extensions of which were the works of Karl Pearson), also coined the term **Eugenics**, the "science" of improving humans through selective breeding. Sir R. A. Fisher, who invented analysis of variance and maximum likelihood techniques, and perhaps more importantly developed the concepts of sampling from populations, degrees of freedom, and his book *Statistical Methods for Research Workers*, is still relevant today.

Eugenics is still with us (click here to access the eugenics-watch website), but has been successfully and completely discredited on scientific grounds many times (click here for Eugenics Archive website). Do keep in mind that the times were different, but it is interesting nevertheless to learn a little about the murky history of statistics and the objectives of some of the very bright people responsible for many of the statistical analyses we use today (see Stephan J. Gould's "*The Mismeasure of Man*" at our Sullivan Library BF 431 G68 1981 or from Amazon.com; Gould, too, may be accused of some bias in his science — see NY Times article based on a PLOS Biology article). Here's an MIT web site with tremendous information about race in science).

Keep in mind also that statisticians were instrumental in showing why Eugenics was unscientific, at best. Here's a link to a non-peer reviewed article.

Questions

1. Find and copy definitions for "big data" and "data mining" from (a) one peer-reviewed, primary source (e.g., search Google Scholar), (b) one peer-reviewed, secondary source (e.g., search Google Scholar), and (c) Wikipedia. From these three sources, write your own definitions for big data processing and data mining.

2.4: Experimental Design and rise of statistics in medical research

Introduction

We were in the fifth month after WHO had declared the Covid-19 pandemic when I last updated this page. If you followed the news at that time you would know of the appeal from some (including the then-President of the United States) for use of hydroxychloroquine, an anti-parasite drug, as a prophylactic or treatment for active Covid-19 infection (cf. Liu et al 2020). The FDA as well as other institutions advised against its use, in part because experimental design concerns were raised for early studies (Kupferschmidt, 2020).

We will spend some time later in the semester with experimental design (Chapter 5), but we can start here. For a number of reasons the **Randomized Control Trial** or RCT — experimental, prospective, double-blind clinical trial with **random selection** of subjects from a **reference population** and **random assignment** of subjects to a treatment group or an appropriate placebo treatment **control group** — is considered the gold standard for producing knowledge (Kaptchuk 2001).

Note:

Experimental control implies researcher imposes conditions to remove possibly confounding effects on the dependent variable — outcome of the experiment. **Placebo**, derived from Latin *placere*, to please (but see Aronson 1999), refers to an inert substance ("sugar pill"), or to a substance with known activity but without effect on the target condition or "wrong indication" (e.g., antibiotics administered for viral infection), given to research subjects in lieu of active treatment. Thus, placebos are examples of treatment controls. The **placebo effect** is improvement of subjects who received the placebo and not active treatment (Pardo-Cabello et al 2022). In contrast, **nocebo effects** are adverse effects attributed to placebo treatment. Under most circumstances, placebo applies to humans only because placebo effects are thought to be product of psychological factors, although mechanisms of action are in dispute. Is sentience necessary for a placebo effect (cf. McMillan 1999)?

Experimental studies imply that the researcher imposed treatments or controls onto subjects. The subjects are followed and outcomes are recorded. Thus, experiments by definition are also **prospective studies** — the outcome is recorded for subjects after some period of time. With a well-designed experiment, the researcher may have evidence to support the claim that, for example, Treatment A causes the outcome.

In contrast, **observational studies** are those in which treatments arise by acts of nature. In both experiments and observational studies, there can be treatment and control groups; the distinction between the types of studies is how assignment of subjects to treatments were affected. Observational studies generally are **retrospective studies** — the outcome has already occurred, the researcher follows up to identify differences among the groups that may account for different outcomes. Examples of observational, retrospective study designs include **cross-sectional** and **case control**; **cohort studies** are prospective studies. Observational studies are discussed further in Chapter 5.4: Clinical trials.

Compared to observational studies, in principle, experiments can establish **cause and effect**. Cause and effect refers to an explanation about relationship between two events or objects. In biology, Ernst Mayr (1904 – 2005) distinguished between two levels of explanation: **proximate** (how) **explanations** and **ultimate** (why) **explanations** (Mayr 1961, cf. Laland et al. 2011). As you know, our mechanism for identifying cause and effect is application of the Scientific Method (Chapter 2.5). Discussions of how to detect cause and effect are provided throughout this book, but emphasized in a few sections (Chapter 16.2 and 16.3).

The principles of good experiments include many steps beyond simply choosing treatments and controls. In Chapter 5 we'll go into more depth, but I wished to list for you some of the key principles of good experimental design. With respect to human-subject research, the researcher needs to protect against many sources of potential bias.

- **Randomization** of subjects assigned to treatment groups controls for individual differences.
- **Placebos** are a means to establish controls in a study so that effects may be attributed to the **active treatment**.
- **Single-blind** implies that the subject does not know what treatment was given. **Double-blind** implies that not only is the subject unaware of the treatment received, but, crucially, neither does the researcher.
- The double-blind design — neither the patient-subject nor the researchers know who received the placebo or the treatment — controls for subtle **biases**.

The experimenter may influence the outcome of the experiment if knowledge about who received the placebo or the new drug; the subject may respond differently with knowledge that they received the placebo and not the new drug. The key intent in this

experimental design is to avoid systematic error, errors in studies that may occur because of our conscious and unconscious beliefs and biases. Placebos are used as treatments because people (and animals!) sometimes get better (or worse) with or without treatment; thus, to be effective, subjects receiving a new drug must get better more frequently than do subjects on placebo. Importantly, the well-designed placebo allows the researcher to gain insight into the mechanism of action by the new drug.

A case to consider

Consider the following experiment (Diener et al 2006; see also Liu et al 2018): subjects who had several migraines per month were treated with acupuncture, sham-acupuncture, or standard treatments including beta blockers, calcium channel blockers, or antiepileptic drugs. After 26 weeks the reductions in reported migraines was compared. The authors reported that there was no difference in numbers of migraines among patients who received the different therapy treatments. The authors conclude that because acupuncture lacks side-effects that may occur with standard therapies, acupuncture may be a good choice for patients seeking relief from migraine.

Another case to consider

Consider the following example. My dad was diagnosed with lung cancer in his early 80s; his left lung showed many spots when imaged and biopsy confirmed the cancer diagnosis. Surgeons removed half of the lung and after five years he was considered cancer free. Why did he develop cancer in the first place? If you immediately think, "He's a smoker," that's not a good explanation: he last smoked tobacco in his early thirties (latency smoking-lung cancer link is about 20 years, Lipfert et al 2019). Tobacco smoking is not the only environmental trigger for lung cancer. Long term exposure to **radon gas**, a naturally occurring, radioactive noble gas, has been linked to lung cancer (EPA). Cancer of the lung in non-smokers is the seventh leading cause of cancer mortality worldwide (Field and Withers 2012). I grew up on Vashon Island, Washington, in a non-smoking home environment. Radon levels on Vashon Island and other areas around Puget Sound are low (source: Washington State Department of Health). Vashon Island is rural, but, as it turns out, within range of a larger copper smelter located in nearby Ruston (Fig. 2.4.1; my home was a 17-km distance from the smelter). The smelter was last in operation in 1986 and was torn down in 1993 (EPA publication number 910R94001). The smelter stack rose more than 500 feet, dispersing smoke laden with heavy metals, notably **arsenic** and **lead**, into the air (Bromenshenk et al. 1985). Over the smelter's 68 years of service, winds carried away the smoke to my island and to other areas known now as the "Ruston-Vashon Island Exposure Pathway" (Kalman et al 1990). Thus, tens of thousands of people were (and continue to be) exposed to the heavy metals deposited into the soils, forming a distinct treatment group (Milham & Strong 1974; Kalman et al 1990; EPA 2000). Is arsenic exposure a plausible mechanism for lung cancer? Workers exposed to arsenic have higher rates of lung cancer (Sullivan 2007, Wei et al 2019). Cultured lung cells exposed to arsenic associated with changes in gene expression (Clancy et al 2012). Coincidentally, two of the family dogs developed and died of cancer, as did one female goat. Perhaps my dad's lung cancer was attributed to long exposure to arsenic (his blood readings for arsenic were in the range of 11 ug/L).

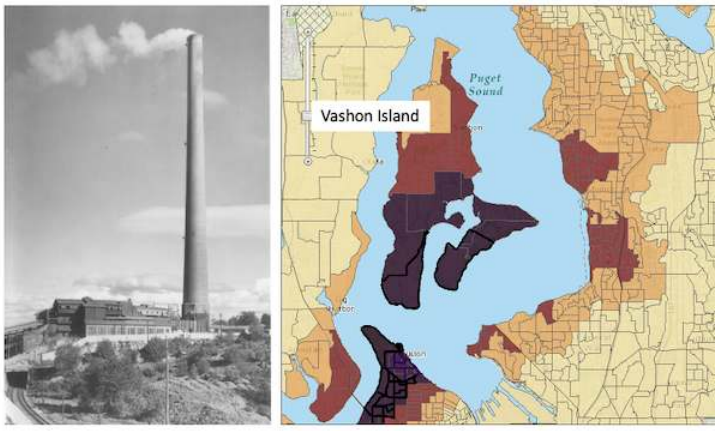


Figure 2.4.1: *Left: ASARCO smelter, Ruston, Washington, image from Department of Ecology, State of Washington.* Direction of smoke from the stack is north, toward Vashon Island. *Right: Heat map of arsenic and lead affected areas. image from kingcounty.gov*. Darker regions correspond to heavier arsenic and lead contamination of soils.

If this scenario seems plausible, I hope you immediately recognize it as a case of **conformation bias** (see Chapter 2.6). Putting aside for a moment the different arsenic species, each with different LD_{50} (the lethal dose needed to kill half the population — see Chapter 20.10), the difficulty ascribing arsenic as a causal agent for my Dad's cancer is that many other exposures happened simultaneously. For example, indoor carpets are a primary source of several volatile organic compounds (Haines et al 2020). Prior

to 1980, carpets may have contained formaldehyde and other known carcinogenic agents. My dad also commuted by car between work and home for decades, this during the early years of the Clean Air Act of Environmental Protection Agency of the United States (it wasn't until 1981 that new cars met EPA emission standards: Clean Air Act timeline here). Thus, all commuters including my Dad were exposed to gasoline combustion emissions, many known to be carcinogenic (Parent et al 2007). Moreover, a limited study by Public Health of Seattle and King County (2001) found that rates of cancer on Vashon between 1980 and 1988 were similar to those in other areas of King County.

Note:

While we "know" tobacco cigarette smoking increases lung cancer risk, and many experiments with animal models convincingly show the link (e.g., Hutt et al 2005), no experiment in the strict sense, i.e., a prospective, randomized control trial, has ever been conducted (hint: it would be unethical; see discussion in Allmark and Tod 2016). Instead, the cumulated evidence from observational studies on exposures of different populations over the years overwhelmingly points to smoking as a leading cause of lung and other cancers.

Questions

1. Was my Dad's lung cancer attributable to his 40-plus year of exposure to soil arsenic (he's a non-smoker)? How should we approach this question?
2. In Diener et al (2006), the authors concluded that because acupuncture lacks side-effects that may occur with standard therapies, acupuncture may be a good choice for patients seeking relief from migraine. Do you agree with the authors?
3. Ethical standards evolve with time. An ongoing debate in research is whether and how placebos are to be used in human subjects research. Placebos are a means to establish controls in a study so that effects may be attributed to the active treatment. The "gold standard" of clinical trials is considered to be the randomized double-blind design — neither the patient-subject nor the researchers know who receives the placebo or the treatment. Following review of the WHO report on *Use of Placebos in Vaccine Trials*, pick one study and evaluate whether or not the decision to use placebos was warranted in your opinion.
4. I searched PUBMED for "double-blind" by decade and found the following results (August 2018) (Table 2.4.1). Open R and/or R Commander and create two variables, then generate a scatter plot. Describe the shape of the relationship between number of publications citing "double-blind" and time (e.g., 1950 – 1959, 1960 – 1969, and so on).

Table 2.4.1. PUBMED results for "double-blind" by decade.

Decade	Publications
1950	60
1960	995
1970	7184
1980	24737
1990	39643
2000	53965
2010	69265

5. Here's one way to enter this data into R. At the R prompt (or in the R Script window of R Commander), create two variables, `Decade` and `Pubs` `Decade <- c(seq(1950, 2020, by=10))` `Pubs <- c(59,995,7161,24728,39670,54011,57043)` Make an XY scatter plot `plot(Decade,Pubs)`
6. Repeat the PUBMED search as above but search for "placebo". Make a table like the one above and provide a scatterplot of your results.
7. Is the concept of a placebo relevant if the subjects in your experiment are yeast cells, not humans?
   - Similarly, if your subjects are yeast cells, how does the concept of performing experiments "blind" apply?
8. Ethical standards change with time. An ongoing debate in research is whether and how placebos are to be used in human subjects research.
   - If placebos are so important, why is their use a concern in clinical trials?

- Following review of the WHO report on Use of Placebos in Vaccine Trials (see Readings at the end of this chapter), pick one study and evaluate whether or not the decision to use placebos was warranted in your opinion.

R notes for question 5:

- `<-` is an assignment operator (`assignOP`); everything to the right of `<-` is assigned to the object named to the left of the `<-` operator. You can instead use `=` in place of `<-` , but because `=` is also used in other contexts besides assignment, a quick look at blogs by data scientists will find a preference to use `<-` for clarity and consistency.
- `c()` "combines" arguments into a vector.
- `seq()` is used to generate a sequence of numbers between a lower and an upper limit; if `by = n` is included, the sequence will be increased by the value n. If omitted, then the sequence is increased by 1.

2.5: Scientific method and where statistics fits

Introduction

The **scientific method** is what makes science a "powerful way of knowing" (Church and Carpenter 2008), and by now, you should be familiar with the outline: hypothesis, experiment, etc. What distinguishes science from other fields of inquiry is that at its best, science accumulates verifiable evidence about our world. I begin with a disclaimer — our introductory textbooks tend to reify the scientific method (cf. Blachowicz 2009). Outside of the classroom and introductory science textbooks, I don't think you'll find much agreement among practicing scientists precisely what the scientific method entails, or whether strict adherence to a list of steps distinguishes what scientists do compared with other professions. For one difference, typical discussions of the scientific method may emphasize the experiment, which probably brings to mind images of a lab coat and test tubes, but should emphasize the critical thinking "tool kit" (Wivagg and Allchin 2002), e.g., model-based inquiry. That said, holding a view that introductory textbooks should present nuances of epistemology seems a big ask.

But we do emphasize experimentation with good reason. In principle it is straightforward to identify what control groups are needed to conduct an experiment in the lab, but what are the control groups for an experiment conducted on the computer? In most cases, one should argue that if an outcome is obtained by random processes, then no additional cause need be addressed. Therefore, the control group for a simulation would be a random process generator.

Disagreements about the scientific method center about how science is really done (e.g., from a social perspective), but also because there appear to be differences in approach in sciences that work on historical questions (physical cosmology, evolutionary biology, geology), and those that conform to the classic experimental approach (chemistry, molecular biology, physics). **Epistemology** is a fascinating area — "How do we know what we know?", "What exactly is science and how is it different from other areas of knowing?", etc. But I will leave you only with the tantalizing suggestion to read more and start with you with a list of readings to start (see Readings at the end of the chapter). This is the stuff of graduate and professional school; we have work to do.

Despite apparent differences between what scientists say they do and how they actually do sciences, there is broad agreement; science as a way of knowing can be characterized by the following steps (National Academy of Sciences 1999).

1. Begin with facts, which are observations confirmed and treated as true
2. Formulate a hypothesis, with emphasis on hypotheses that are testable statements about relationships observed about the natural world
3. Given the possible outcomes, state predictions derived from the hypothesis.
4. Make observations of perform an experiment designed to test the hypothesis.
5. Analyze the data from the experiment.
6. Evaluate the results of the experiment against the predictions.
7. Repeat.

This deceptively simple list hides much work to be done. Hypotheses are not "educated guesses," where "educated guess" implies an idea about how some phenomenon is likely the correct explanation because of the skill or knowledge of the person making the guess. Good hypotheses make possible experimental tests whose results can be used to rule out alternative explanations, *sensu* Platt's "strong-inference".

Hypothetico-deductive reasoning

I mention this disclaimer about the common (and reassuring) textbook discourse on the scientific method to suggest that, if you have not already reached this point in your career, it is time to move past the cookbook approach to thinking about what it means to do science in practice. There is induction and deduction and probabilistic thinking that must be grappled with, all emphasizing efforts by the individual, and yet science if it is to make any progress must ultimately be a communal activity (Varmus 2009). In particular, to the extent a researcher consistently applies hypothetico-deductive reasoning, or as Platt (1964) called it, "strong-inference," then good science can happen (see Fudge 2014 for an update). Strong inference according to Platt implies that researchers should follow three steps (after Fudge 2014):

1. develop alternative hypotheses
2. think of a crucial experiment that can exclude one or more hypotheses
3. perform the experiment and obtain a clean result.

Then, beginning with step 1, repeat the procedure to refine the possibilities that remain.

The list of elements of the scientific method no where point to the crucial role of scientists engaged in an active community of scientists fits in. However, we can can quickly suppose that every step of the scientific method can involve input from others to help shape, improve, and indeed carry out the activities needed to practice science. What scientists share is critical thinking and the tools of statistics provides a common language.

It's all about the probability of a particular event

Platt (1964) wrote that some of the observations we make are puzzling or hard to explain, which implies our understanding is incomplete. We then proceed to ask questions about why the observations are different from our expectations, and we speculate about how the outcome comes about. Thus, we have to consider the **probability**, or chance, of a particular outcome (event) compared to other possible outcomes. Statistics is about analyzing the probability of outcomes. But there is a twist; there are two distinct, but complementary, approaches to statistics. Most of the statistics you have been taught so far comes from the **frequentist approach**. That is, how often (frequency) will we get the kind of results we observe given the hypothesis? Another approach can be termed **Bayesian**, where the question is, how likely is the hypothesis to be correct given the data? These two approaches view the data obtained from the same experiment differently. A frequentist views the data as random — repeat the experiment and the results will differ — but the hypothesis is fixed (it is either true or it is not). A Bayesian, on the other hand, views the hypothesis as random with a probability of being true somewhere between zero and 100 percent, and the data are fixed.

Which are you, a frequentist or a Bayesian? Consider the following from xkcd.com no. 1132 (Fig. 2.5.1).

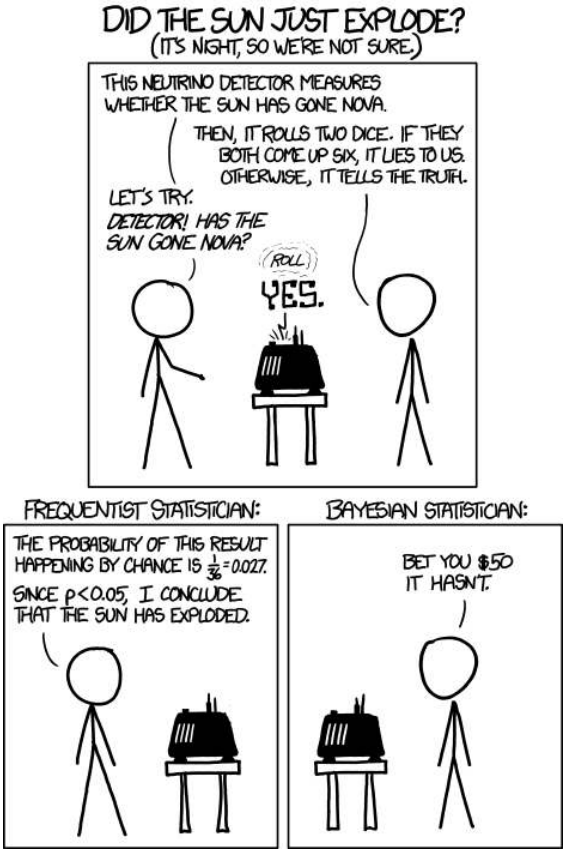


Figure 2.5.1: "Frequentists vs Bayesians," xkcd.com no. 1132.

The Bayesian approach makes sense when there exists **prior knowledge** — the Sun came up this morning, as it did the day before, and the day before that, etc. New data are assessed against what we already know. The Frequentist approach, despite philosophical shortcomings, works for analyzing experiments where prior knowledge is lacking about outcomes. Because much of biomedical research is based on the Frequentist approach, most of our efforts, too, will utilize that approach.

A statistical translation of the Scientific Method

Given our frequentist perspective, we can outline our Scientific Method as follows.

1. Formulate a Null and an Alternate hypothesis.

2. State predictions from the Null and an Alternate hypotheses.
3. Design an experiment or observation.
4. Analyze the data from the experiment or observation.
5. Interpret the experiment or observation.
6. Evaluate the predictions from the Null and an Alternate hypotheses.
7. Accept (provisionally) or reject (provisionally) the Null hypothesis.
8. Evaluate model fit and robustness
9. Repeat.

In practice, there is more to statistics than these 8 (or 9) steps, but this does provide the outline of what statistical analyses are about. There are nuances to how a Frequentist or a Bayesian views statistical analyses as evidence for or against a conclusion (Goodman 1999a, 1999b). We need to distinguish between when data are acquired in the process; data may come before or after the hypotheses are stated. In epistemology, hypotheses are either *a priori* or *a posterori* and we need to add these to our discussion of Scientific Method. The terms are Latin, translated apparently as "from the earlier" and "from the later", respectively

Note from the lists described as "the" scientific method how hypothesis comes first. Folks who think and write about how and why we know what we know — the discipline is called epistemology — tell us we are generally on more solid footing when we design experiments with specific intent, specific and testable hypotheses. The branch of statistics concerned with experimental design provides rich context for many practical aspects of how to implement experiments — in other words, how to follow the Scientific Method, even if turns out there isn't just one universal definition of Scientific Methods out there.

When we say "experiment," the hypothesis came first

More commonly in statistics, the phrases **planned**, and therefore *a priori*, and **unplanned** or *a posterori*, comparisons are referenced. In practice, biologists design experiments and make observations accordingly to test one or more hypotheses, but they may also address additional hypotheses after the fact, especially if the experiment generates a lot of data.

You may have also heard of the phrase **data mining**. Data mining is loosely defined, but mostly refers to sets of protocols and procedures to extract patterns from large data sets stored in databases. Google apparently does lots of data mining, as do many other businesses that obtain large amounts of data. Unplanned comparisons include any data mining protocols, no matter how sophisticated the language sounds (feature selection, classification tree). Data mining is not consistent with classic experiments; it's different than Step 6 (Step 8 in the second list) listed above because no new data or experiment is carried out.

Can you get away with coming up with and testing new hypotheses from data gathered from an experiment designed to test a different hypothesis? Yes, and of course, the process can be quite profitable for Google. However, you should proceed with caution and restraint. If you are not careful about how you write it up — you will probably be called on it by a reviewer of your work. Is this *a posterori* approach still science? Of course, yes! I'd even go so far as to say that when one studies real systems, you can't limit yourself to only *planned* hypotheses and testing. At least, one designs experiments and carries out those tests but then also uses current data to generate new ideas. Science is also exploratory — you may not *design* an experiment at all, but through observations, you probably will develop *testable* hypotheses! We'll return to these concepts soon. What you must be aware of in any unplanned comparison or data mining sojourn is the possibility of committing a **data dredging** (aka **p-hacking**) sin — searching through data to come up with misleading, but statistically "significant" results (Ioannidis 2007; Stefan and Schönbrodt 2023).

Testing of unplanned *a posterori* hypotheses is a real concern in science. On the one hand, those who think about how we learn about the world and make sense of it have stated emphatically that the best way we know is to follow the scientific method — and that begins with hypotheses followed by designed experiments to test claims derived from those a priori hypotheses. We (teachers, textbook author) continue to teach science as the act of individuals toiling away in the forest or in the lab, pursuing sets of questions that may involve the collection of measurements on dozens to a few hundred subjects. While this type of science is still in practice, there is no doubt that big science involving many people is more common and, perhaps, better at generating new knowledge (Wuchty et al 2007). One result of "big science" is to generate a lot of data, the very essence now of big data, and there must be room for testing of new hypotheses gathered on data sets. This is the essence of the argument for the ENCODE project (ENCODE Project Consortium 2012), which generates lots of genetic data on the human genome using common techniques and makes the data publicly available — new research can be conducted on old data.

Bottom-up, top-down

Consider the pronouncements (almost daily, it seems!) about the discovery of a new gene for some disease, process, or behavior in humans. Often times, although not always, these "discoveries" are not duplicated by other research groups. Why not? Well, for one, the phrase "gene for" is a dubious short-hand for what is usually a more complex causation. But from our statistical perspective it is problematic because the search for genes is really an *a posterori* exercise — one begins with phenotypic differences (some have the disease, others do not) and some genetic information (SNPs, DNA sequencing) and then proceeds to see if there are any differences in the genetic material at hand between the two groups. This approach, the Genome Wide Association Study, or GWAS, would be termed "**top-down**" — begin with the phenotypic differences and search for genetic differences between those that do and those who do not have the condition. Sampling is an issue (are the unaffected subjects a random sample from the entire population?), but the problem also is one of logical design — the hypothesis is made after the fact — a statistical difference between the groups is attributed to a particular genetic difference.

Case-control subjects, where patients with the condition are matched with other individuals who do not have the condition, but match in other ways (e.g., age, income, etc.), are enrolled in such studies because there can be no random assignment of subjects to treatment. Case-control subjects are selected such that affected and unaffected individuals are matched by characteristics in as many meaningful ways as possible (e.g., age, gender, income, etc.). Because many tests are conducted in GWAS studies, i.e., is there a difference between the control and affected group for the first gene, the second gene, and so on up to the number of genes on the microarray chip (10,000 or more genes), the chance that any particular association is a **false positive** is high, discussed further when we cover Risk Analysis in Chapter 7.

How to interpret a test result

You are a medical doctor reading the results of a test for three-month average glycated hemoglobin, **A1C**, levels for your patient. For example, A1C above 6% is considered strong marker of diabetes. A marker, yes, but not the same as a guarantee that a person has diabetes. A **false positive** (FP) is the case where a test result is positive, but the subject in fact does not have the condition. False positive is equal to the ratio of

$$False\ positive\ rate = \frac{FP}{FP+TN}$$

where FP is the number of false positive readings and TN refers to the **true negative**, the number of those who in fact do not have the condition.

In Chapter 7 we will spend some time on risk analysis. To introduce this important subject in biostatistics, we'll begin with an example. A study of 15,934 subjects without diagnosed diabetes found that 3.8%, or 605 individuals, had elevated A1C levels, which translates to about 7.1 million U.S. adults not yet diagnosed with diabetes (Selvin et al 2011). About 90% of these individuals also had fasting glucose levels greater than 100 mg/dL, i.e., diabetic. This is clearly a good test. However, note that 10% of the 605 individuals had elevated A1C, but fasting glucose levels less than 100 mg/dl, i.e., did not meet the diagnostic of type 2 diabetes. Thus, the false positive rate is about 0.4%, or nearly 30,000 adults with elevated A1C without diabetes.

Working through frequencies can be challenging, so applying a natural number approach helps. Having just read through the frequencies and percentages, now look at how they translate to a **probability tree** (Fig. 2.5.2). Start by utilizing a per-capita rate standard: for proportions in the 10 – 20% range, a standard 1000 persons works well (Fig. 2.5.2).

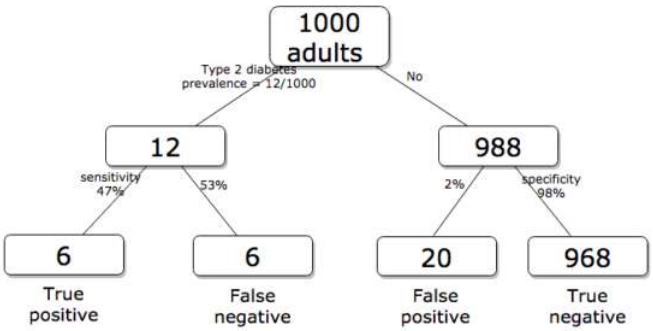


Figure 2.5.2: Probability tree diagram with prevalence of type 2 diabetes and sensitivity, specificity of A1C test, data from CDC and Selvin et al 2011. Tree drawn with free diagrams.net app.

How many individuals are expected to have A1C above 6%? Twenty-six total positives, of which only 23% are true positives. An important lesson is that if the prevalence of a condition is low, then any diagnostic test with high sensitivity necessarily will identify many false positives.

False discovery rate: claims must be stringently evaluated

The genome scientists involved in GWAS studies generally are aware of false positives, also termed **false discovery rate** (FDR), conduct statistical corrections to account for false positives (e.g., Brzyski et al 2016), and generally are cautious in their interpretation. But not always. Studies of associations between autism and environment come to mind (e.g., Waldman et al 2008), and recent developments in the direct-to-consumer genetic testing market also suggest that the limitations of these kinds of studies are not always represented.

There is a broader concern about the **reliability of research**, and the debate about how to improve reliability comes from a call to understand how to do statistics better and, more importantly to understand how statistics are to be used in making claims from statistical results (Ioannidis 2007; but see Goodman and Greenland 2007). A key element of scientific work is that findings are repeatable: results from one group should apply to other groups. One dictate to improve reproducibility — increase number of subjects in studies — is obvious, but given the cost of GWAS, currently an unreasonable demand.

In some cases, biologists already have a particular gene in mind, whose function is more or less known, and then the exercise follows the Scientific Method listing much more directly. This **bottom-up** approach leads to a straightforward, testable genetic hypotheses: a specific difference in genomic sequence predicts a difference in phenotype outcome. A good example is the identification of more than 100 different single nucleotide mutations, called single nucleotide polymorphisms or SNPs, in the CFTR gene of patients known to have cystic fibrosis disease (Castellani 2013). As of August 2018, the number of known pathogenic or likely pathogenic SNPs is now listed at 440 in the SNP database; to put this in some context, there are more than 40,000 reported single nucleotide polymorphisms for the CFTR gene (this number includes SNP duplicates).

Conclusions

Epistemology, the theory of how knowledge is acquired, is a complicated business – what I want you to appreciate now is that planned and unplanned comparisons affect interpretation of your statistical results, how the difference is likely also to affect the reproducibility of your work. In the language or clinical trials and experimental design, planned and unplanned accompany **prospective** and **retrospective** studies (Chapter 5). A prospective study in the case of GWAS means genetic differences among individuals are known at the start of the study, and phenotypic differences arise naturally during the course of the study. Knowing the strengths and limitations of, for example, a planned retrospective study is at the heart of experimental design (Chapter 5).

Questions

1. Follow links to and read papers by Platt (1964) and Cleland (2001) to answer the following questions.
   - What is the problem with a scientist coming up with only one hypothesis for his or her research?
   - What did Platt mean by “strong inference” and how did he recommend this be accomplished?
   - According to Cleland (2001) how do experimental sciences and historical sciences differ in how they handle the asymmetry of causation?
2. Above we described how GWAS studies are generally top-down; what would constitute a bottom-up approach to GWAS?
3. In your own words, provide pro and con points of view on data mining of large data sets.
4. What would be the harm of a false positive in a GWAS study of prostate cancer (review in Benafif et al 2018)?

2.6: Statistical reasoning

Introduction

So far in the course we have talked about **statistical reasoning**. Statistical reasoning as part of the **critical thinking** tool kit. What this should mean for you in practice is that we will be developing and enhancing your critical thinking skill set. Here, we'll define our terms and place statistical reasoning in context.

Critical thinking

What do I mean by critical thinking? You'll find various definitions and discussions, but the Wikipedia entry on the subject seems as good a start as anything else I have heard or read:

> "Critical thinking is a way of deciding whether a claim is always true, sometimes true, partly true, or false" (5 August 2013, http://en.Wikipedia.org/wiki/Critical_thinking).

Which raises the question, what is truth (false)? And what about the frequency qualifiers "always", "sometimes", or "partly"? This hints at one of the strengths of science: "Truth in science … is never final, and what is accepted as fact today may be modified or even discarded tomorrow" (p. 2, National Academy of Science 1999). Scientists extend and even correct the work of their predecessors.

Critical thinking has many definitions, but they coalesce broadly as the search for **rational justification** — arguments and decisions based on reason not emotion — given sets of **facts**. We'll define facts as assertions about the world that are **verifiably true**. Related to facts, we define **evidence** as collection of facts used to infer an assertion or belief is **objectively true**. Objectivity holds independently of the observer. In practice, critical thinking can be improved by identifying and developing certain skill sets, from deductive reasoning to use of a statistical toolkit, although skills alone do not necessarily lead to sound critical thinking (Bailin 2002). Emphasis on the importance of statistical reasoning, for example in the **Evidence-Based Medicine** (EBM) approach, is now well established. EBM is an expansive concept, but includes the philosophy that decisions in medicine and health care should be based on evidence (results) from well designed and executed research. More generally, much has been written on the subject. My purpose here is to sell you just a bit that you will learn more than formulas and statistical tests in this course; if we do this right you will improve your critical thinking skills.

And the tool kit for critical thinking? Science has the **scientific method** to offer. In turn, by adopting statistical reasoning approaches you will be more cognizant of whether or not you are indeed engaged in critical thinking about your work.

A working checklist for critical evaluation of a project:

- "All sides" of the problem are fairly represented
- Authority is not synonymous with empirical truth: trust, but verify
- Correlation (association) may or may not infer causation
- Do conflicts of interest compromise conclusions?

In other words, identify and rigorously consider assumptions made to reach conclusions, evidence given in support of conclusions, and considerations of bias held by the researcher. The "all sides" part requires judgement — not all sides of a research question are equally valid; the point about "fair" representation requires that opposing arguments are actually held by others and not simply a strawman characterization of of a position held by no one.

Rigid adherence to a "how to do science checklist" is likely to be wanting, and thus practicing and successful scientists show more flexibility. And, the list is certainly incomplete because it presumes observations and data collection are from a reliable, unbiased source. Before proceeding with how statistics informs our critical thinking in science, let's illustrate bias.

Bias

Bias is defined as any tendency that may hamper your ability to answer a question without prejudice (Pannucci and Wilkins 2010). Although we tend to think of ourselves as highly rational, psychologists have documented and named numerous, and in some cases, highly specific kinds of cognitive biases we may have about "**real world data**" that may impair our judgment about results of experiments. At its best, application of the scientific method is our best tool kit to help protect our conclusions against bias (Nuzzo 2015), by proper experimental design and possible study biases. A couple of comics from xkcd illustrate.

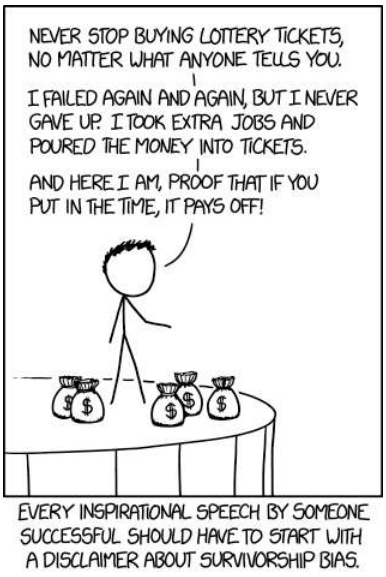


Figure 2.6.1: "Survivorship bias," https://www.xkcd.com/1827/.

Figure 2.6.2: "Selection bias," https://xkcd.com/2618/.

Wikipedia lists dozens of named biases which may influence our ability to interpret science:

- **confirmation bias**, e.g., given extraneous information about ancestry or sex, forensic investigators were more likely to assign ancestry or sex of skeletal evidence to a group that confirmed the contextual information (Nakhaeizadeh et al 2014).
- **congruence bias**, a type of confirmation bias, the tendency to test only a favorite or initial hypothesis over consideration of alternative hypotheses.
- **observer-expectancy bias**, e.g., randomized controlled trials of acupuncture often find equivalent responses to real and placebo acupuncture despite both appearing superior to no treatment (Colagiuri and Smith 2012).
- **selection bias**, the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved, e.g., mitotic counts (Cree et al 2021) and cell differentials and Figure 2.6.2. Selection bias, where the study population is not well-defined or if is, the method by which subjects are recruited favors one kind of subject over another.
- **surveillance bias**, e.g., association between myocarditis and COVID-19 vaccines may be due to increased focus to identify myocarditis (Husby et al 2021).
- **survivorship bias**, focusing on subjects that passed a selection process as if they represent an entire group, e.g., from ecology: rare species more likely to go extinct than abundant species (Lockwood 2003) and Figure 1.

Bias in research may occur at any level, from study design to analysis and publication of results. Bias leads to **systematic errors** that may favor one outcome over others. Clearly then, bias is something one wants to avoid in science, and most scientists would probably agree strongly that bias is unacceptable. However, bias can creep into projects in many forms. Particularly pernicious causes of bias are conflicts of interest, with the notable case of a 1990's set of trials in gene therapy (Wilson 2010). Biased research may be more harmful to science than deliberate **misconduct**, at least in part because there are, or should be, mechanisms to detect fraud (e.g., Marusic et al 2007). If errors in data analysis or management are rare, we can trust the results; if errors are common, systematic, or deliberate falsification, then any conclusions drawn from such work deserves retraction at the very least (cf. discussion in Baggerly and Coombs 2009; see StatCheck project, Nuijten et al 12016).

Bias in research

Bias is not limited to obvious profit-seeking or even defense of one's ideas or reputation. Bias can enter research in subtle ways. Prior to 2000, many post-menopausal women could expect to be placed on hormone replacement therapy (HRT) to manage post-menopausal symptoms and to reduce likelihood of osteoporosis later in life. Moreover, HRT was believed based on research to show protective benefits against heart disease (e.g., Nabulsi et al 1993; Grodstein et al 2000). However, other prospective studies that were designed to reduce many sources of bias in patient recruitment (Women's Health Initiative Study Group 1998), found the opposite: HRT may increase risk of heart disease as well as other diseases (Rossouw et al 2002). Since the 2002 Women's Health Initiative study report, the costs and benefits of HRT have been furiously debated (Lukes 2008).

The purpose of this chapter, however, is not to exhaustively review sources of bias. Instead, we leave you with a partial list of kinds of bias in clinical trials (from a review by Pannucci and Wilkins 2010).

- **Bias during study design**, e.g., use of subjective measures or poorly designed questionnaires.
- **Interviewer bias**, for example, if the interviewer is aware of the subject's condition
- **Chronology bias**, where the control subjects may not be observed in the same time frame as the treatment subjects
- **Citation** (reporting) **bias**, where negative results are not reported
- **Confounding**, where results may be due to factors not properly controlled that also affect the outcome. An example provided was link between income and health status, which would be confounded by access to health care.

These are but a few of the kinds of bias that even proper research may be influenced by. Solutions are to randomize, to double-blind, and to avoid reporting bias. Controls of sources of bias increase the validity of the research. How experimental design may control for bias is discussed further in Chapter 5.3.

A case study

Stories we hear from the news or from our friends about health or the environment are anecdotal. A basic rule in critical thinking is to distinguish between argument that is built on anecdotal evidence and argument based on scientific evidence. Let's evaluate the following real-world example.

Airborne® is a leading dietary supplement. Here's what their website has to say about the product.

"Scientific research confirms that Airborne proprietary formulation with 13 vitamins and minerals plus a blend of health-promoting herbs, does indeed enhance immunity" (5 August 2013, http://www.airbornehealth.com/our-story).

This next statement appears in smaller text near the bottom of the page.

"These statements have not been evaluated by the Food and Drug Administration. These products are not intended to diagnose, treat, cure or prevent disease" (5 August 2013, http://www.airbornehealth.com/our-story).

So, in short, how do you evaluate Airborne? The claim is that "scientific research confirms that Airborne … enhance immunity," and yet, there is a statement that follows that suggests that the claim has not been evaluated by the agency legally responsible in the USA for determining the efficacy of medicine and treatments. Are not these in contradiction? In the USA, supplement manufacturers need not seek FDA approval. In fact, the U.S. Congress, in its passing of the The Dietary Supplement Health and Education Act of 1994 (DSHEA), specified what needs to be reported to the FDA; statements of "enhancing immunity" are lawful, but do not require the supplement manufacture to show evidence that their product provides this support, only that ingredients in the supplement may do so.

Returning to the statements from Airborne Health, how may these statements be interpreted?

1. peer-reviewed articles testing Airborne are available
2. that experiments were conducted without bias
3. and that to "enhance immunity" is something one could measure

Let's look at #3 first. A search of PubMed (December 2013) for the phrase "enhance immunity" returned 9171 hits. I didn't look through the more than 9000 articles, but a quick look shows that the term is indeed used in research, but carries a wide range of interpretations. For one paper "enhance immune" response was the result of genetic modifications to T-cells and how they responded to a virus (PMID:24324159).

Adding "vitamin" to the search resulted in just 158 hits in PubMed; none of these are about Airborne (see claim #1 and #2). However, to be fair, go back and look at their statement — notice they are not saying *exactly* that anyone has studied Airborne (otherwise they'd probably list the papers), just that vitamins and minerals and herbs that are in Airborne have been studied.

Indeed, vitamins A, C, E and minerals like zinc and selenium that are in Airborne are among the 158 papers. To be additionally fair to Airborne, there are indeed many papers investigating benefits (harm) of supplemental vitamins and minerals — and again, being charitable here, what we can say is that results are at best inconsistent (Bjelakovic et al 2014; Comerford 2013).

The manufacturer of Airborne ran into some trouble with the FTC and ended up paying over $20 million US dollars to settle the lawsuit. The one study, not peer-reviewed, of Airborne's effectiveness cited during these proceedings apparently was sponsored by Airborne's manufacturer. Airborne was acquired by Schiff Vitamins March 2012 (Business Wire April 2 2012). Click here for some additional reading about Airborne.

About peer-review

Evaluation of claims made, research reported, and ideas in science are generally expected to be subject to peer-review by knowledgeable persons in the field. Effective peer-review is an important "gaketekeeper" in science (Siler et al 2015, Tennant 2018). Proper peer review gives legitimacy to research (Siler et al 2015). The traditional peer review system is a closed system, which includes elements controlled by journal editors and is based on reviewer identify remaining anonymous. The closed peer-review system is not the only option: many have argued for an open review process (e.g., Pöschl 2012, Ho et al 2013).

Does the Airborne story meet your definition of bias?

Airborne made health claims about their product; these claims were found by the federal government to be misleading but does not fit into our discussion of bias. Instead, the Airborne story fits more with critical thinking: I think it is rather a casual position to say that we should, and do, think critically of news sources or the stories our friends tell us about health.

But bias in science occurs too, and can be very difficult to control or completely eliminate.

Learning statistical reasoning will help develop critical thinking skills

In addition to considering sources of bias, i.e., can you trust the messenger and thinking about the sets of facts offered to bolster a claim, it helps to evaluate each fact to see if:

- The author is really talking about correlation and has not in fact provided evidence of causation
  - vs. experiment, which tests "cause" manipulation of X, does Y change?
- Measurement is appropriate
  - Direct or indirect
- Unmeasured important (co)variables exist

Questions

1. The xkcd comic in Figure 2.6.1 calls it survivorship bias. Provide a definition for this kind of bias and describe the effects on conclusions if it is not accounted for.
2. Doctors are encouraged to engage in "evidence based medicine," that is, decisions about care should be based on data. How may confirmatory bias (from the patient's point of view, from the doctor's point of view) influence a discussion of "facts" when deciding on a course of medication therapy — for example, between a well-established medication with many known side-effects and a new, recently approved medicine reported to be as effective as the standard treatment?

2.7: Chapter 2 References and Suggested Readings

Allmark, P, & Tod, AM (2016). Ethical challenges in conducting clinical research in lung cancer. *Translational Lung Cancer Research*, *5*(3), 219. https://doi.org/10.21037%2Ftlcr.2016.03.04.

Aronson, J (1999). Please, please me. *Bmj*, *318*(7185), 716. https://doi.org/10.1136/bmj.318.7185.716.

Association of American Medical Colleges (2020). What's on the MCAT2020 exam? Retrieved from https://students-residents.aamc.org/applying-medical-school/article/mcat-2015-sirs-skill4/.

Baggerly KA, Coombes KR (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics* 3(4):1309-1334. https://doi.org/10.1214/09-AOAS291.

Bailin S (2002). Critical thinking and science. *Science & Education* 11:361-375. https://doi.org/10.1023/A:1016042608621.

Benafif, S., Kote-Jarai, Z., & Eeles, R. A. (2018). A Review of Prostate Cancer Genome-Wide Association Studies (GWAS). *Cancer Epidemiology and Prevention Biomarkers* (8), 845-857. https://doi.org/10.1158/1055-9965.epi-16-1046.

Bjelakovic G, Nikolova D, Gluud C. (2014). Antioxidant supplements and mortality. *Current Opinion in Clinical Nutrition and Metababolic Care* 17:40-4. https://doi.org/10.1097/mco.0000000000000009.

Blachowicz, J. (2009). How science textbooks treat scientific method: A philosopher's perspective. *The British Journal for the Philosophy of Science*.

Brieger K, Hardin J (2012). Medicine, statistics, and Education: the inextricable link. *Chance* 25(3):31-34.

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, *72*(1), 2-10.

Bromenshenk, J. J., Carlson, S. R., Simpson, J. C., & Thomas, J. M. (1985). Pollution monitoring of Puget Sound with honey bees. *Science*, 227(4687), 632-634.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., & Sabatti, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics*, *205*(1), 61-75.

Camões, J. (2016). *Data at work: Best practices for creating effective charts and information graphics in Microsoft Excel*. New Riders.

Castellani, C. (2013). CFTR2: How will it help care? *Paediatric Respiratory Reviews* 14 (supplement 1): 2-5.

CDC. (2021, July 21). *Chloroquine or hydroxychloroquine*. National Institutes of Health. Retrieved August 20, 2021, from https://www.covid19treatmentguidelines.nih.gov/therapies/antiviral-therapy/chloroquine-or-hydroxychloroquine-and-or-azithromycin/ .

Clancy, H. A., Sun, H., Passantino, L., Kluz, T., Muñoz, A., Zavadil, J., & Costa, M. (2012). Gene expression changes in human lung cells exposed to arsenic, chromium, nickel or vanadium indicate the first steps in cancer. *Metallomics*, *4*(8), 784-793.

Cleland, C. (2001). Historical science, experimental science, and the scientific method. *Geology* 29:987-990.

Colagiuri, B., & Smith, C. A. (2012). A systematic review of the effect of expectancy on treatment responses to acupuncture. *Evidence-based complementary and alternative medicine*, *2012*.

Comerford, K.B. (2013). Recent Developments in Multivitamin/Mineral Research. *Advances in Nutrition* 4(6):644-656.

Cree, I. A., Tan, P. H., Travis, W. D., Wesseling, P., Yagi, Y., White, V. A., … & Scolyer, R. A. (2021). Counting mitoses: SI (ze) matters!. *Modern Pathology*, *34*(9), 1651-1657.

Diener, H. C., Kronfeld, K., Boewing, G., Lungenhausen, M., Maier, C., Molsberger, A., … & GERAC Migraine Study Group. (2006). Efficacy of acupuncture for the prophylaxis of migraine: a multicentre randomised controlled clinical trial. *The Lancet Neurology* 5(4), 310-316.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414), 57.

Fudge, D. S. (2014). Fifty years of JR Platt's strong inference. *Journal of Experimental Biology*, 217(8), 1202-1204.

Goodman, S.N. (1999a). Toward evidence-based medical statistics. 1. The P value fallacy. *Annals of Internal Medicine* 130:995-1004.

Goodman, S.N. (1999b). Toward evidence-based medical statistics. 2. The Bayes factor. *Annals of Internal Medicine* 130:1005-1113.

Grodstein, F.; Manson, J. E.; Colditz, G. A.; Willett, W. C.; Speizer, F. E. & Stampfer, M. J. (2000) A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Annals of Internal Medicine* 133:933-941.

Haines, S. R., Adams, R. I., Boor, B. E., Bruton, T. A., Downey, J., Ferro, A. R., … & Dannemiller, K. C. (2020). Ten questions concerning the implications of carpet on indoor chemistry and microbiology. *Building and Environment*, *170*, 106589.

Ho, R. C. M., Mak, K. K., Tao, R., Lu, Y., Day, J. R., & Pan, F. (2013). Views on the peer review system of biomedical journals: an online survey of academics from high-ranking universities. *BMC medical research methodology*, 13(1), 1-15.

Husby, A., Hansen, J. V., Fosbøl, E., Thiesson, E. M., Madsen, M., Thomsen, R. W., … & Hviid, A. (2021). SARS-CoV-2 vaccination and myocarditis or myopericarditis: population based cohort study. *bmj*, *375*.

Hutt, J. A., Vuillemenot, B. R., Barr, E. B., Grimes, M. J., Hahn, F. F., Hobbs, C. H., … & Belinsky, S. A. (2005). Life-span inhalation exposure to mainstream cigarette smoke induces lung cancer in B6C3F1 mice through genetic and epigenetic pathways. *Carcinogenesis*, *26*(11), 1999-2009.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* 2(8): e124. doi:10.1371/journal.pmed.0020124.

Kalman, D. A., Hughes, J., van Belle, G., Burbacher, T., Bolgiano, D., Coble, K., … & Polissar, L. (1990). The effect of variable environmental arsenic contamination on urinary concentrations of arsenic species. *Environmental health perspectives*, 89, 145-151.

Kaptchuk, T. (2003). Effect of interpretative bias on research evidence. *British Medical Journal* 326(7404): 1453–1455.

Kaptchuk, T.J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology* 54:541-549.

Laland, K. N., Sterelny, K., Odling-Smee, J., Hoppitt, W., & Uller, T. (2011). Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful?. *Science*, *334*(6062), 1512-1516.

Lipfert, F. W., & Wyzga, R. E. (2019). Longitudinal relationships between lung cancer mortality rates, smoking, and ambient air quality: a comprehensive review and analysis. *Critical Reviews in Toxicology*, *49*(9), 790-818.

Liu, J., Cao, R., Xu, M., Wang, X., Zhang, H., Hu, H., … & Wang, M. (2020). Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. *Cell discovery*, *6*(1), 1-4.

Liu, L., Zhao, L. P., Zhang, C. S., Zeng, L., Wang, K., Zhao, J., … & Li, B. (2018). Acupuncture as prophylaxis for chronic migraine: a protocol for a single-blinded, double-dummy randomised controlled trial. *BMJ open*, *8*(5), e020653.

Lockwood, R. (2003). Abundance not linked to survival across the end-Cretaceous mass extinction: patterns in North American bivalves. *Proceedings of the National Academy of Sciences*, *100*(5), 2478-2482.

Luckhoff, C. (2021). Congruence Bias. In *Decision Making in Emergency Medicine* (pp. 89-96). Springer, Singapore.

Lukes, A. (2008). Evolving issues in the clinical and managed care settings on the management of menopause following the women's health initiative. *The Journal of Managed Care & Specialty Pharmacy* 14:7-13.

Kroese, D. P., Brereton, T., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(6), 386-392.

Kupferschmidt, K. (2020). Three big studies dim hopes that hydroxychloroquine can treat or prevent COVID-19. *Science*, *368*, 6496.

McMillan, F. D. (1999). The placebo effect in animals. *Journal of the American Veterinary Medical Association* 215(7):992-999.

Marusic, A., Katavic, V., & Marusic, M. (2007). Role of editors and journals in detecting and preventing scientific misconduct: strengths, weaknesses, opportunities, and threats. *Medicine and Law*, 26, 545.

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*(3489), 1501-1506.

McCullough, B.D., Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics & Data Analysis* 49(4):1244-1252.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

Milham Jr, S., & Strong, T. (1974). Human arsenic exposure in relation to a copper smelter. *Environmental Research*, 7(2), 176-182.

Muenchen R. A. (2019). *The popularity of data science software*. Retrieved from http://r4stats.com/articles/popularity/ .

Nabulsi, A. A. and Folsom, A. R. and White, A. and Patsch, W. and Heiss, G. and Wu, K. K. and Szklo, M. (1993) Association of hormone-replacement therapy with various cardiovascular risk factors in postmenopausal women. The Atherosclerosis Risk in Communities Study Investigators. *New England Journal of Medicine* 328:1069-1075.

Nakhaeizadeh, S., Dror, I. E., & Morgan, R. M. (2014). Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. *Science & Justice*, *54*(3), 208-214.

National Academy of Sciences Steering Committee on Science and Creationism. (1999). *Science and Creationism: A view from the National Academy of Sciences*, 2nd edition, National Academy Press.

National Research Council Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century. (2003). *Bio210: Transforming Undergraduate Education for Future Research Biologists*. Washington, DC: National Academy.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, 48(4), 1205-1226.

Nuzzo, R. (2015). How scientists fool themselves–and how they can stop. *Nature*, *526*(7572), 182-185.

Pannucci, C.G., and Wilkins, E.G. (2010). Identifying and Avoiding Bias in Research. *Plastic Reconstruction Surgery* 126(2): 619–625.

Pardo-Cabello, A. J., Manzano-Gamero, V., & Puche-Cañas, E. (2022). Placebo: a brief updated review. *Naunyn-schmiedeberg's Archives of Pharmacology*, *395*(11), 1343-1356.

Parent, M. E., Rousseau, M. C., Boffetta, P., Cohen, A., & Siemiatycki, J. (2007). Exposure to diesel and gasoline engine emissions and the risk of lung cancer. *American journal of epidemiology*, *165*(1), 53-62.

Platt, J.R. (1964). Strong inference. *Science* 146:347-353 (link).

Pöschl, U. (2012). Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. Frontiers in computational neuroscience, 6, 33.

Raju, C. K. (2011). Probability in ancient India. In *Philosophy of Statistics* (pp. 1175-1195). North-Holland.

Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, Kotchen JM, Ockene J.(2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* 288(3):321-33.

Sackett, D. L. (1979). Bias in analytic research. In *The case-control study: consensus and controversy*. *Journal of Chronic Diseases* 32(1-2):51-63.

Sackett, D. L., Rosenberg W. M., Gray J. A., Haynes R. B., Richardson W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal* 312(7023): 71–72.

Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Holt, New York.

Science and Creationism: A view from the National Academy of Science (1999), by Steering Committee on Science and Creationism.

Seattle and King County Public Health Report (2001). Review of available data on types of cancer related to arsenic exposure: Vashon-Maury Island, Washington state and Washington state counties 1980-1998.

Selvin, E., Steffes, M. W., Gregg, E., Brancati, F. L., & Coresh, J. (2011). Performance of A1C for the classification and prediction of diabetes. *Diabetes care*, 34(1), 84-89.

Selvin, E., Zhu, H., & Brancati, F. L. (2009). Elevated A1C in adults without a history of diabetes in the US. *Diabetes Care*, 32(5), 828-833.

Shin, A. (2008). Airborne coughs up millions to settle suits. Friday, August 15, 2008 *Washington Post*, https://www.washingtonpost.com/wp-dyn/content/article/2008/08/14/AR2008081403142.html .

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, 112(2), 360-365.

Stefan, M.I., Gutlerner, J.L., Born, R.T., Springer, M. (2015). The Quantitative Methods Boot Camp: Teaching Quantitative Thinking and Computing Skills to Graduate Students in the Life Sciences. *PLoS Comput Biol* 11(4): e1004208.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346. https://doi.org/10.1098/rsos.220346 .

Stricker, D. (2008). BrightStat.com: Free statistics online. *Computer Methods and Programs in Biomedicine*, 92, 135-143.

Sullivan, M. (2007). Contested science and exposed workers: ASARCO and the occupational standard for inorganic arsenic. *Public health reports*, 122(4), 541-547.

Tennant, J. P. (2018). The state of the art in peer review. FEMS Microbiology letters, 365(19), fny204 https://doi.org/10.1093/femsle/fny204 .

The Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials* 19:61-109.

United States Environmental Protection Agency, Region 10 (2000). *First Five Year Review Report for Ruston/North Tacoma Superfund Site Ruston and Tacoma, Washington.*

Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO reports*, *13*(4), 291-296.

Waldman, M., Nicholson, S., Adilov, N., Williams, J. (2008). Autism Prevalence and Precipitation Rates in California, Oregon, and Washington Counties. *Arch Pediatr Adolesc Med.* 2008;162(11):1026-103.

Wilson, R. F. (2010). The death of Jesse Gelsinger: new evidence of the influence of money and prestige in human research. *American journal of law & medicine* 36(2-3), 295-325.

Wivagg, D., & Allchin, D. (2002). The dogma of "the" scientific method. *The American Biology Teacher*, *64*(9), 645-646.

Wuchty S, Jones B, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316:1036-1039.

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome biology*, *17*(1), 1-3.

CHAPTER OVERVIEW

3: Exploring Data

Introduction

After an experiment has been completed, the collection of observations needs to be summarized or described, a process now referred to as **data exploration**. How many mice grew tumors in the treated versus control group? Did all mice respond to treatment? What is the typical cost of a new home in Hilo, HI? How large of a difference is there between the most expensive home and the middle? These questions require simple, basic summary statistics or **descriptive statistics** and informative **data visualizations**.

In this section we introduce two aspects of **summary statistics** expected in any report of a **data set**. Summary statistics provide a brief overview of relevant characteristics of observations in the data set to provide the reader to get a quick but meaningful look at the data set. Data set refers to a collection of data, usually a collection of related, ordered observations, measurements, and related information.

Summary statistics vary according to the needs of the reporting vehicle, but may generally address two characteristics of the data set

1. Central tendency or the description of the middle of the data.
2. Dispersion, the variability around the middle of the data set.

We introduce statistical graphics, a specialized kind of data visualization, in Chapter 4 – How to Report Statistics. Statistical graphs are utilized to describe data sets, but also to communicate statistical inference, which we address formally beginning in Chapter 7 – Probability and Risk Analysis.

Most of you have been asked at some point to calculate the average or the standard deviation. We will provide these again, but with additional statistics. Textbooks may present **calculator formulas** — nothing wrong with them, although we have to watch **significant figures**. But computer statistical packages generally do not use these formulations — that's why I present the formulas throughout the book to help define the statistical concept, not as a way to necessarily calculate the statistic by hand calculation. I haven't investigated this last point in any systematic way, but, because of access by scientists to increasingly powerful computers since the 1980s, I doubt anyone in the business of data analysis in the biological sciences has much use of the hand calculator.

> **Note to BI311 students:**
>
> On homework, quizzes and exams, you may be asked to calculate these descriptive statistics. I will provide you with formulas that illuminate the definitions of the statistics rather than enhance their computation.

3.1: Data types

Introduction

Data? Data refers to collections of facts, information, or statistics about an object. Data are either quantitative (numbers) or qualitative (observed properties that cannot be summarized by numbers). Data are **measured** and analyzed for research or reports to be used as evidence in support or against some hypothesis or for some other decision making arena (medicine, policy). Measurement implies a systematic effort to assign a numerical value to the thing that is measured; **measurement units** are standard quantities used to describe the same kinds of things. Examples of measurement units include kilograms (mass), meter (length), liter (volume), and Celsius (temperature).

Data also implies a means to code or structure information so that it can be analyzed. **Raw data** refers to unprocessed collection of information about an object, which then needs to go through **data processing** in order to be useful in the next steps. If you look more closely, you'll see that considerable effort is made to standardize data formats for analytical purposes. Good examples of such standards are available in clinical research and genomics.

In statistics, we recognize data which belongs to either of two **data types**: quantitative or qualitative. We will return to data types repeatedly throughout our statistics journey — knowing which type you directs you to the types of statistical tests that are available to you. In brief, **quantitative data types** implies estimation of parameters about a population, hence, this data type points the user towards use of parametric statistics; **qualitative data types** do not lead to estimates of parameters, but provide counting of observations in categories.

Quantitative data

Discrete: countable or meristic, example: five *Conus* shells (Fig 3.1.1)

Figure 3.1.1: Five *Conus* shells, example of discrete data type.

Interval: example: degrees Celsius (Fig. 3.1.2)

Figure 3.1.2: Thermometer showing office temperature at 23.1 Celsius, example of interval data type.

Ratio, true zero, examples: body mass, capillary blood glucose reading (Fig. 3.1.3), degrees Kelvin, relative humidity (Fig. 3.1.4).

Figure 3.1.3: Blood glucose reading, 122 mg/dL.

Figure 3.1.4: Hygrometer showing office humidity at 65 percent, example of ratio data type.

Qualitative data

Binomial, yes/no, example: a person either has the condition or they do not; hydrangea petals may or may not be blue (Fig. 3.1.5).

Figure 3.1.5: Flowers are blue or they are not, example of binomial data type.

Nominal, example: names of species. Wolves and dogs are members of *Canis lupus and Canis familiaris*, respectively; house cats are not (Fig. 3.1.6).

Figure 3.1.6: Cats are neither dogs or wolves, example of nominal data type.

Note:

Identifying variables, or **id numbers**, are unique identification numbers or other for each record (individual) in the data set. These variables are categorical, nominal data type. Examples of id numbers include Social Security numbers, student identification numbers, driver's license numbers, etc. Note that id numbers would only rarely be considered objects of study because they are typically assigned by researchers to subjects and not properties of subjects. Exceptions may include testing for impacts of anonymization procedures (for example, see Koll et al 2022).

Ordinal, ranked, example: Likert scale:

- Strongly disagree
- Disagree
- No opinion
- Agree
- Strongly agree

Although common practice, caution is warranted when converting Likert categories into numerical scale, for example, Strongly agree = 4, Strongly disagree = -4, and so on. Because it is ordinal, the difference between 4 and -4 can't be calculated as the difference because it is ranked, not the numerical scale.

Biologists should know their data types before proceeding with an experiment.

Examples to try

In R, load the data set `diabetic` (`survival` package, which is loaded as part of R Commander), then view the variables.

For more about R data sets, see Part 6: Working with an included data set in Mike's Workbook for Biostatistics

R code

```
data(diabetic, package="survival")
```

In R Commander (Fig. 3.1.7):

> **Rcmdr: Data → Data in packages → Read data set from an attached package…** Double click `survival`, the list of data sets should appear in the right-hand panel. Select `diabetic`, then click OK button.

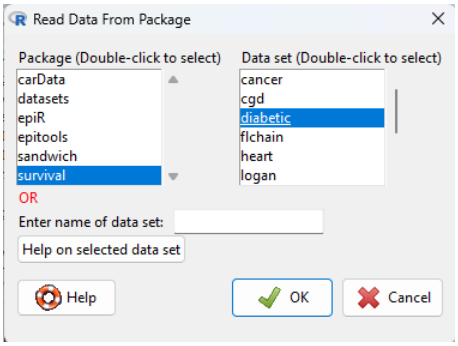


Figure 3.1.7: Screenshot of Read Data from Package menu in R Commander.

View the data by clicking on Rcmdr's View data set button, or, better, submit the following command in R:

```
head(diabetic)
```

R output:

```
  id laser age   eye trt risk  time status
1  5 argon  28  left   0    9 46.23      0
2  5 argon  28 right   1    9 46.23      0
3 14 xenon  12  left   1    8 42.50      0
4 14 xenon  12 right   0    6 31.30      1
5 16 xenon  9   left   1   11 42.27      0
6 16 xenon  9  right   0   11 42.27      0
```

The command `head()` displays by default the first six rows of a data frame.

It's a good idea to read up on the data set. Data sets included with R packages often provide a help page. Submit the following command in R to load the help page.

```
help(diabetic)
```

The data set was subjects with high risk diabetic retinopathy; "each patient had one eye randomized to laser treatment and the other eye received no treatment."

What are the data types for the variables? I'll give you the a couple to start. The first column with entries 1 – 6 is called the **index variable**; it's row 1, row 2, etc. of the data set and technically is not a data set variable (since its assignment is arbitrary) — R adds this for you. Next, the variable labeled `id` — clearly we see numbers, so we might think meristic, but because these are labels for the subjects, the proper data type is nominal! Try identifying the data types and example units of measurement for the rest on your own, then open the hidden text immediately below to see the best answers.

Answers to Examples to Try

laser: binomial, there were two types (xenon or argon)

age: ratio, years

eye: binomial

trt: binomial, no treatment (0) or laser (1)

risk: ordinal

time: ratio, time to event, number of months

status: binomial

Questions

Assign the data type and examples of units of measurement for each kind of measurement.

1. Darts tossed, Distance from center.
2. Shells, width, length.
3. InfraRed temperature device readings.
4. Body weight.
5. Lung volume.
6. Tomato color morphs (green, yellow).
7. Tomato root length, stem length.
8. Systolic blood pressure.
9. Blood arsenic levels.
10. Body Mass Index.
11. Body Mass Index scale, for example NIH: underweight, normal, overweight, obese.

3.2: Measures of central tendency

Introduction

For a sample of observations we can begin the summary by identifying the "typical" value. Various statistics are used to describe the middle and collectively these are referred to as measures of **central tendency**. The **mean**, the **median**, and the **mode** are the most common measures of central tendency. In the situation in which we work with data from a population **census**, we would calculate **population descriptive statistics** — not **inferential statistics**; because we more often work with samples from populations, we report **sample descriptive statistics**.

First we review the familiar arithmetic mean and introduce the weighted mean. Next we introduce "other means," which you may not be as familiar with.

Means

There are several means beyond the simple arithmetic average or mean. Here we review a few. We use this topic also to start introducing standard notation we will use throughout the book.

The **population mean**, μ (pronounced "mu"), is given by

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

where X_i is an observation on the i^{th} individual,

N is the size of the population, and

$\sum$ or "sigma" instructs you to add up the X values from $i = 1$ (the first observation) to $i = N$ (the last observation).

The **sample mean**, $\bar{X}$, pronounced "eks bar," of a collection of observations is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where n is the size of the sample.

> **Note:**
>
> **Parameters** (aka **random variables**) get Greek letters and **sample variables** get Roman letters. See Chapter 3.4.

Weighted arithmetic mean

In some cases you may have several samples from the same population. If the sample sizes are the same, you can calculate the average of averages without any fuss — just take all of the sample means and add them up, then divide by the total number of samples. If the sample sizes differ, then you need to weight (W) each sample mean by its sample size. Simply divide each sample mean by its appropriate sample size, then add all of these up. That is the weighted average.

More generally, we can write

$$\bar{X}_W = \frac{\sum_{i=1}^{n} W_i \cdot X_i}{\sum_{i=1}^{n} W_i}$$

For example, consider a variable containing the following observations

Table 3.2.1. A sample of observations.

Observation	Frequency
4	4
5	2
6	3
7	2

Observation	Frequency
8	3
15	1

The observation “4” was observed four times; the observation “5” was observed twice, and so on for a total of 15 observations.

What is the arithmetic mean of these 15 observations? To solve this, well, you have a couple of choices. You could copy the numbers down as often as they appear and then calculate the mean in the usual way.

Other “means”

The arithmetic average (illustrated above) is not the only way to estimate the mean.

The **trimmed mean**, also called the **truncated mean**, is a useful approach when data is widely dispersed — data spread away from the middle (see Chapter 3.3). Thus, the trimmed mean will be less influenced compared to the arithmetic mean by **outlier data points**, i.e., data far from other data points in the set.

You would use the trimmed mean to describe the middle of a data set in which a plot shows most of the values are clumped together around a middle – and yet you see a few values that are much smaller or much greater. A specified percentage of the smallest and largest values are removed from the data set and then the simple arithmetic mean is calculated for the trimmed data set. For example, given a data set of daily rainfall for different cities, you might wish to remove the driest 5% and wettest 5% of the days in order to better compare the rainfall trends for the cities.

Calculating the trimmed mean is straight-forward in R: use the same built-in function, `mean()` , but add some options.

Note:

This is a good point to remind you **how to get help with R commands**. Do you recall how to get help in R?

At the R prompt type

```
help(mean)
```

or

```
?mean
```

The R Documentation page for `mean()` will pop up (assuming you allowed R to install help pages as html). Figure 3.2.1 shows a screenshot of a portion of the help page for `mean()`

mean {base} R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

x	An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects. Complex vectors are allowed for `trim = 0`, only.
trim	the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.
...	further arguments passed to or from other methods.

Figure 3.2.1: A portion of the R help page about the function mean.

From the help page (Fig. 3.2.1) we can see that we can specify a trimmed mean by adding options to the `mean(x, ...)` command. For our `x` variable defined above, get the trimmed mean after 25% of the data are removed.

```
mean(x, trim=0.25)
```

```
[1] 6
```

Note from the help page that the only required option you need to feed the mean command is the name of the variable, in this case, "x" (it can be, of course, any name provided the data are attached). In this case we removed 25% of the values – 12.5% of the smallest values and 12.5% of the largest values – that's also called the interquartile mean.

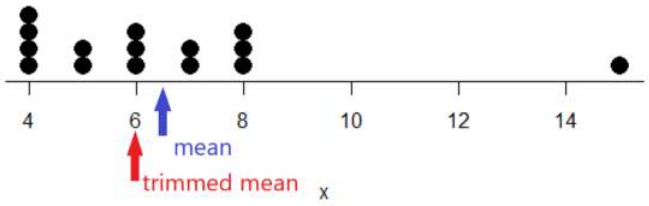


Figure 3.2.2: Dot plot of our x variable with locations of the mean (blue) and the trimmed mean (red). The `Dotplot(x)` function in package `RcmdrMisc` was used in `Rcmdr` to make this graphic. Arrows were added by hand. `Dotplot()` example code presented in Chapter 3.4.

If we recalculate a trimmed mean after dropping 10% of the points, or even 40% of the points, we get the same mean value of 6. The trimmed mean is an example of a **robust estimator**; it's resistant to the influence of outliers.

Another useful descriptor of the middle is the **geometric mean**. The geometric mean is useful for calculating the average of ratios. Geometric mean would be used when you want to compare central tendency for different variables, each differing in scale. For example, gene expression results, reported as fold-changes, for different genes often shows tremendous differences among genes and would be best described by logarithmic scale, not arithmetic scale. Geometric mean expression values would be better choice for central tendency. Other examples are found in economics: for example, calculating compound interest or interest. The geometric mean applies whenever the scale is multiplicative and not additive.

The geometric mean is given by the equation

$$gm = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

The geometric mean (gm) is equivalent to **log-transforming** your data, then calculating the arithmetic mean, and transforming the result back (with the **antilog** exponent.) As you recall, for our simple data set the arithmetic mean was 6.2. The geometric mean for this data was 5.977. Taking the natural log for each of the values from our simple data set, then calculating the arithmetic mean we have 1.788.

The antilog of this value is

```
exp(1.788)
```

```
[1] 5.977
```

Another frequently encountered mean is the **harmonic mean**, which is defined by the equation \[H = \frac{n}{\sum_{i=1}^{n} \frac{1}{\bar{X_{i}}} \nonumber\]

Harmonic mean is appropriate for averaging rates. For example, what is the average speed traveled if you travel 30 miles per hour (mph) between point A and B, then on the return trip, your speed was 40 mph? If you think (30 + 40)/2 = 35 mph, then this would be incorrect — after all, the distance covered has not changed, just the time. The harmonic mean returns 34.2 mph. Let

```
y = c(30,40)
```

The harmonic mean returns 34.2 mph (see below, "How to calculate these other means")

Both harmonic and geometric means apply for values greater than zero.

How to calculate these "other" means

In Microsoft Excel, calculate geometric mean via the function `GEOMEAN()` ; calculate harmonic mean via the function `HARMEAN()` .

The base R (and Rcmdr) doesn't have built in functions for these, although you could download and install some R packages which do (e.g., package `psych` , `geometric.mean(variable)` , `harmonic.mean(variable)`). It is quicker to just to calculate these by submitting a snippet of code into the script window

For geometric mean of variable "x" at the R prompt type

```
exp(mean(log(x)))
```

For harmonic mean of variable "x" at the R prompt type

```
1/mean(1/x)
```

where is the base of the natural logarithm, **Euler's number**, and log is the **natural logarithm** (in R, to get log to other bases you can use `log10` for **base 10 logarithm** or `log2` for **base 2 logarithm**, or `log(x, base = n )` for any base n of the variable `x` , and variable is the name of the variable you wish to do the calculations on.

R code: Do try on your own!

Here's some numbers to try your hand. For example, create a variable containing a few numbers, any numbers, and write it to the variable named `z`

```
z = c(3,4,6,7,9,11,4)
```

Now, calculate the arithmetic mean, the geometric mean, and the harmonic mean for the variable `z` . Using the values shown above for `z` , you should get Table 3.2.2.

Table 3.2.2. Comparison of different means for `z` .

arithmetic mean	6.285714
geometric mean	5.716903
harmonic mean	5.204936

Try three more. In R (or a R Commander script window), create three new variables.

```
varA = c(3,3,3,3)
```

```
varB = c(1,2,3,12)
```

```
varC = c(-3,0,1,3)
```

Now, calculate the arithmetic mean, geometric mean, and harmonic mean for each variable.

For the simple arithmetic mean

```
mean(varA)
```

For the geometric mean, use the formula above

```
exp(mean(log(varA)))
```

For the harmonic mean, use the formula above

```
1/mean(1/varA)
```

What did you get?

Other measures of central tendency

Median

The median, which lacks an accepted notation — we'll go with $Med(x)$, divides a set of observed numbers into two equal halves. Half the observations are above the median, half of the observations are below the median. Arrange data from lowest to highest, take the middle measurement:

$$\frac{1}{2}\, observations\ ranked < median < \frac{1}{2}\, observations\ ranked$$

For an odd number of measurements, the median is the middle value. For an even number of measurements, the median is the average of the 2 middle values. Or more succintly, we have

$$Med(x) = \begin{cases} X\left[\frac{n+1}{2}\right] & \text{if } n \text{ is odd} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n+1}{2}}}{2} & \text{if } n \text{ is even} \end{cases}$$

To get the median in R, type at the R prompt

```
median(variable)
```

and of course, replace `variable` with the name of the variable containing the numbers. For our `x` variable created earlier, the function median returns in R

```
[1] 6
```

[Note that the median for `x` was the same as the trimmed mean for `x`, which is consistent with with our view that the trimmed mean is a robust estimator of the middle of a data set.

Mode

Mode is another way to express the middle and it refers to the most frequent occurring measurement. Use of mode makes most sense for discrete or countable numbers. For a normal distribution, the mean, median and mode will be the same value. Note that a data set may have more than one mode. For example, what is the mode for the variable we created earlier?

```
x = c(4,4,4,4,5,5,6,6,6,7,7,8,8,8,15)
```

For this small data set we see that "4" is the most frequent with a count of four occurrences in the set.

Mode would seem like a straightforward function in R. However, it turns out there is not a mode function in the base package.

A little explanation is in order. In R, typing `mode` at the R prompt like so

```
mode(x)
```

returns

```
[1] "numeric"
```

Not the answer we were expecting. In R, mode command is used to tell you what the mode (i.e., way or manner in which some task is accomplished) of storage is for the variables.

In order to get the statistical mode we want, we either hunt down a package that contains mode estimation (e.g., install the package `modeest`, use the `mfv` function), or we can write a little code.

Note:

Although the `modeest` package is available from the typical repositories, the `genefilter` dependency required by `modeest` is available through Bioconductor. Bioconductor is an R repository dedicated to R packages for genomic data analysis.

A quick Google search found a number of answers at stackoverflow.com (e.g., question 2547402). The simplest response was to use the `names` and `max` commands like so:

```
temp = table(as.vector(x))
```

```
names (temp)[temp==max(temp)]
```

Why the median may be a better middle than the mean

Comparing the two measures of central tendency can tell you without plotting how your data are distributed about the middle. Sample distributions are discussed in Chapter 6.

- When the distribution of the data is symmetric or normally distributed (discussed in Chapter 6.7) then the mean and the median will be about the same value
- When data are **right-skewed** (a few large values), then the mean will be *greater* than the median.
- When the data are **left-skewed** (a few small values), then the mean will be *less* than the median.

Here's an illustration (Fig. 3.2.3). I sampled 100 points from a **random normal distribution** with mean zero and standard deviation one, and another 100 points from a log-normal distribution also with mean zero and standard deviation one. In Figure

3.2.3, the **histograms** (see Chapter 4.2 – Histograms) of both data sets are shown, along with **summary statistics** (see Note below). Means are indicated with red arrows and medians are indicated with blue arrows.

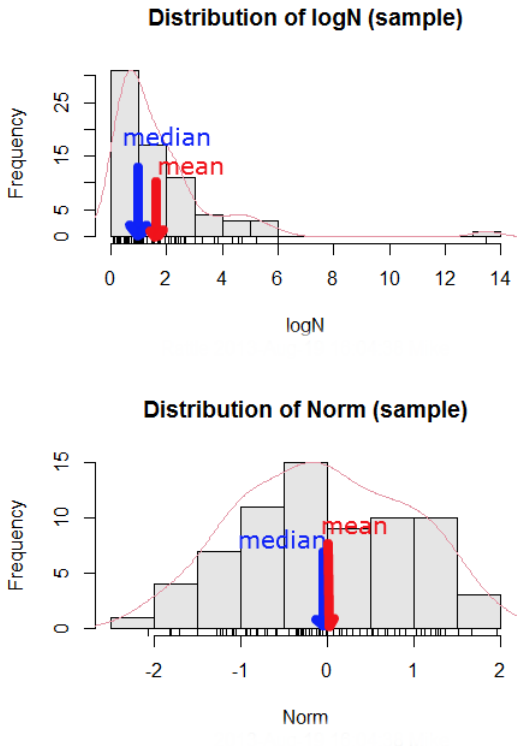


Figure 3.2.3: Normal and lognormal distributions with mean (labeled in red) and median (labeled in blue) noted for comparison.

So, the median is a better descriptor of the central tendency of a sample distribution when the distribution is NOT normally distributed.

> **Note:**
>
> "Summary statistics" refers to reporting of one or more descriptive statistics on a data set. The mean, median, standard deviation, range are common reported statistics. R Commander provides a menu to select from descriptive statistics, returning a table of the estimates. **Rcmdr: Statistics → Summaries → Numerical summaries…**

Scaling and transformation of data

Sometimes it is useful to **standardize** your data so that the variables all have the same scale. One algorithm for standardization is called **normalization**. Normalization implies that you correct the data so that data has a mean, μ, of zero, and a standard deviation, σ, of 1 (unit variance). There are several ways to standardize, each with strengths and limitations. To normalize we use the **Z-score** equation (see Chapter 6.7 for other uses of Z score).

$$Z = \frac{X_i - \mu}{\sigma}$$

where X_i is each observation in your data set.

Normalization will make **outliers**, the few points in a data set that are noticeably different from the central tendency of the rest of the data, smaller and less influential. When you normalize multiple sets of data, then each will have the same mean (zero) and variance (unit variance), but the ranges will differ. An example of this is the simple product moment correlation — by standardizing you change the variances for the different variables to have the same unit variance.

As we will see later in class it is also useful to expand or contract the variability of the data or to change the shape of the distribution (if the data is not normally distributed). For example, if you compare individuals of a population for many morphological traits (e.g. body size, growth rate), the spread of points (called a distribution) will look more like a Poisson

distribution (not symmetrical about the mean, a few individuals may be much larger…). This is partly due to the way in which morphological traits are measured. We normally measure body size on a linear scale (inches or centimeters). However, body size is affected by physiological processes that are more related to volume. Therefore, the more appropriate scale of measurement is on a log scale. We can **transform** the data measured on a linear scale to a log scale. For morphological traits this can produce a distribution that is normally distributed (bell shape). There are many more statistical procedures for data that is normally distributed than there are statistical procedures for Poisson distributions or any other type of distribution. Additional discussion about data transformation is introduced in Chapter 13.3.

You can always **uncode** or **unstandardize** your data after performing the statistical procedures and return to the original scale. In fact, when reporting descriptive statistics you should report the untransformed, uncoded data. Moreover, you will find it useful to report means adjusted for other variables (e.g., from ANOVA or regression); if the ANOVA or regression equations are performed on transformed or coded data you would want to back calculate to the original scale after applying the ANOVA or regression adjustments. This advise will make more sense after we've discussed ANOVA (Chapter 12) and linear regression (Chapter 17).

R operators

The `names` command can be used to retrieve the names contained in the variable (if text types) or to set the names of the observations, which is what we are using it for here. We set the numbers to text names "4", "5", etc. then find the maximum count of named items in the temp table. The double equals operator (==) is used to tell R to find the object that is "equal to" something we specify, in this case, the max value (R Language Definition 2014). Table 3.2.3 shows common operators in R.

Table 3.2.3. Common arithmetic* and comparison** operators

?	help
+	plus
–	minus
*	multiply
/	divide
:	series
>	greater than
<	less than
>=	greater than or equal to
<=	less than or equal to
=	left assignment
<-	left assignment
==	equal to

* To list and get help with use of arithmetic operators enter at the R prompt

```
help(Arithmetic)
```

** To list and get help with use of comparison operators enter at the R prompt

```
help(Comparison)
```

The R package `modeest` has a number of algorithms for calculating the mode, depending on the kind of data you are working with. After installing the package and its dependencies, type at the R prompt

```
require(modeest)
mfv(x)
[1] 4
```

Creating objects in R

Everything in R is an object (Chambers 2008). Create the variable in R by assigning the vector `x`, either directly at the R prompt or in a script window (Rcmdr, RStudio), like so

```
x = c(4,4,4,4,5,5,6,6,6,7,7,8,8,8,15)
```

The function `c()`, which stands for combine, is used to combine the set of numbers into the object, `x`. For small sets like this you may find it convenient to enter the values one by one and let R store it into the vector for you. Use the `scan()` function and your keyboard. Careful! Make sure that you remember to assign the results from scan to a vector.

I'll create the object `tryScan` just to distinguish it from `x`, although I will enter the same values. Until R receives an interrupt signal from you, it will prompt you to enter numbers one row at a time. When you've reached the end, use the keyboard combination `Ctrl+q` (`Command + q` on Macs) to interrupt keyboard input.

```
tryScan = scan()
1: 4
2: 4
3: 4
4: 4
5: 5
6: 5
7: 6
8: 6
9: 6
10: 7
11: 7
12: 8
13: 8
14: 8
15: 15
Read 15 items
```

```
tryScan
 [1] 4 4 4 4 5 5 6 6 6 7 7 8 8 8 15
```

The function `tryScan` is a very useful command, with many options, and can be used for more than keyboard entry. For example, you can paste a column of numbers from your spreadsheet using your computer's clipboard.

R code calculate central tendency

Once we have the vector `x`, calculate the mean by entering at the R prompt

```
mean(x)
```

and you should get the answer of `6.466667`

And of course, you don't type in the R prompt `>`, right?

Or, for the better option, create two variables, one containing the list of observed numbers and the second that contains the frequency for each observed number in the series. You would then use the command for weighted mean.

```
y = c(4,5,6,7,8,15)
```

```
w = c(4/15, 1/15, 3/15, 2/15, 4/15, 1/15)
```

Note — you can check that the frequencies sum to 1 by using the sum command like so:

```
sum(w)
```

For the weighted mean, the command is

```
weighted.mean(y,w)
```

and the answer returned is `6.466667` , the same as before.

Questions

1. Find the help page in R for the median function. How does the function handle missing values?
2. For a simple data set like the following `y <- c(1,1,3,6)` you should now be able to calculate, by hand, the
   - mean
   - median
   - mode
3. If the observations for a ratio scale variable are normally (symmetrically) distributed, which statistic of central tendency is best (e.g., less sensitive to outlier values)?
4. In the `names()` command, what do you think the result will be if you replace `max` in the command with `min` ?
5. If data are right skewed, what will be the order of the mean, median, and mode?
6. Calculate the sample mean, median, and mode for the following data sets:
   • Basal 5 hour fasting plasma glucose-to-insulin ratio of four inbred strains of mice,
   ```
   x <- c(44, 100, 105, 107) #(data from Berglund et al 2008)
   ```
   • Height in inches of mothers,
   ```
   mom <- c(67, 66.5, 64, 58.5, 68, 66.5) #(data from GaltonFamilies in R package
   HistData)
   ```
   and fathers,
   ```
   dad <- c(78.5, 75.5, 75, 75, 74, 74) #(data from GaltonFamilies in R package
   HistData)
   ```
   • Carbon dioxide (CO_2) readings from Mauna Loa for the month of December for demi-decade 1960 – 2020
   ```
   years <-c (1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010, 2015,
   2020) #obviously, do not calculate statistics on years; you can use to make a plot
   co2 <- c(316.19, 319.42, 325.13, 330.62, 338.29, 346.12, 354.41, 360.82, 396.83,
   380.31, 389.99, 402.06, 414.26) #data from Dr. Pieter Tans, NOAA/GML
   (gml.noaa.gov/ccgg/trends/) and Dr. Ralph Keeling, Scripps Institution of
   Oceanography (scrippsco2.ucsd.edu/)
   ```
   • Body mass of *Rhinella marina* (formerly *Bufo marinus*, Fig. 3.2.4),
   ```
   bufo <- c(71.3, 71.4, 74.1, 85.4, 85.4, 86.6, 97.4, 99.6, 107, 115.7, 135.7, 156.2)
   ```

Figure 3.2.4: Female *Rhinella marina* (formerly *Bufo marinus*), Chaminade University campus. Body length 23.5 cm.

3.3: Measures of dispersion

Introduction

The statistics of data exploration involve calculating estimates of the middle, central tendency, and the variability, or dispersion, about the middle. Statistics about the middle were presented in the previous section, Chapter 3.2. Statistics about **measures of dispersion**, and how to calculate them in R, are presented in this page. Use of Z score to standardize or normalize scores is introduced. Statistical bias is also introduced.

Describing the middle of the data gives your reader a sense of what was the typical observation for that variable. Next, your reader will want to know something about the variation about the middle value — what was the smallest value observed? What was the largest value observed? Were data widely scattered or clumped about the middle?

Measures of dispersion or **variability** are descriptive statistics used to answer these kinds of questions. Variability statistics are very important and we will use them throughout the course. A key descriptive statement of your data, how variable?

> Note:
>
> **Data** is plural = a bunch of points or values; **datum** is the singular and rarely used.

Examine the two figures below (Fig. 3.3.1 and Fig. 3.3.2): the two **sample frequency distributions** (put data into several groups, from low to high, and membership is counted for each group) have similar central tendencies (mean), but they have different degrees of variability (standard deviation).

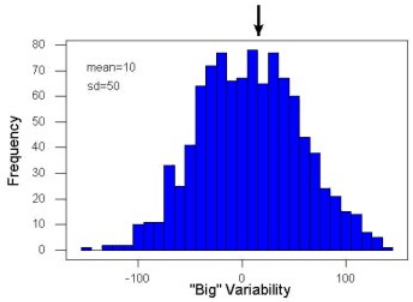


Figure 3.3.1: A histogram which displays a sampling of data with a mean of 10 (arrow marks the spot) and standard deviation (sd) of 50 units.

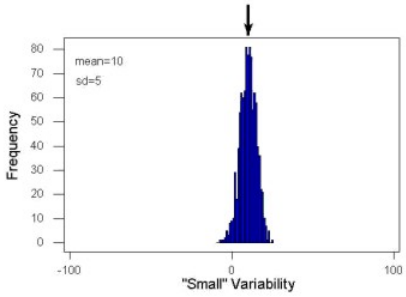


Figure 3.3.2: A histogram which displays a sampling of data with the same mean of 10 (arrow marks the spot) displayed in Fig. 3.3.1, but with a smaller standard deviation (sd) of 5 units.

Clearly, knowing something about the middle of a data set is only part of the required information we need when we explore a data set; we need measures of dispersion as well. Provided certain assumptions about the frequency of observations hold, estimates of the middle (e.g., mean, median) and the dispersion (e.g., standard deviation) are adequate to describe the properties of any observed data set.

For measures of dispersion or variability, the most common statistics are: Range, Mean Deviation, Variance, Standard Deviation, and Coefficient of Variation.

The range

The range is reported as one number, the difference between maximum and minimum values.

$$range = X_{max} - X_{min}$$

Arrange the data from low to high, subtract the minimum, X_{min}, from the maximum, X_{max}, value and that's the range.

For example, the minimum height of a collection of fern trees might be 1 meter whereas the maximum height might be 2.2 meters. Therefore, the range is 1.2 meters (= 2.2 – 1).

The range is useful, but may be misleading (as index of true variability in population) if there is one or more exceptional "outlier" (one individual that has an exceptionally large or small value). I often just report X_{min} and X_{max}. That's the way R does it, too.

If you recall, we used the following as an example data set.

```
x = c(4,4,4,4,5,5,6,6,6,7,7,8,8,8,15)
```

The command for range in R is simply `range()` .

```
range(x)
[1] 4 15
```

There is apparently no universally accepted symbol for "range," so we report the statistic as $range = 15 - 4 = 11$.

You may run into an **interquartile range**, which is abbreviated as **IQR**. It is a trimmed range, which means, like the trimmed mean, it will be robust to outlier observations. To calculate this statistic, divide the data into fourths, termed **quartiles**. For our example variable, x, we can get the quartiles in R:

```
quantile(x)
  0%  25%  50%  75% 100%
 4.0  4.5  6.0  7.5 15.0
```

Thus, we see that 25% of the observations are 4.5 (or less), the second quartile is the median, and 75% of the observations are less than 7.5. The IQR is the difference between the first quartile, Q1, and the third quartile, Q3.

The command for obtaining the IQR in R is simply `IQR()` . Yes, you have to capitalize "IQR" — R is case sensitive; that means IqR is not the same as IQR or iqr.

```
IQR(x)
[1] 3.5
```

and we report the statistic as *IQR* = 3.5. Thus, 75% of the observations are within 3.5 points of the median. The IQR is used in box plots (see Chapter 4). Quartiles are special cases of the more general **quantile**; quantiles divide up the range of values into groups with equal probabilities. For quartiles, four groups at 25% intervals. Other common quantiles include **deciles** (ten equal groups, 10% intervals) and percentiles (100 groups, 1% intervals).

The mean deviation

$$sample\ mean\ deviation = \frac{\sum_{i=1}^{n} \left| X_i - \bar{X} \right|}{n}$$

Subtract each observation from the sample mean; each $X_i - \bar{X}$ is called a ***deviate***: some observations will be positive (greater than $\bar{X}$) and some will be negative (less than $\bar{X}$).

Take the absolute value of the deviation and then add up the absolute values of the deviations from the mean. At the end, divide by the sample size, n. Large values for this statistic imply that much of the data is spread out, far from the mean. Small values in turn imply that each observation is close to the mean.

Note:

The mean deviation will always be positive, which is why we take the absolute. By taking the absolute value of each deviate, then the sum is greater than zero. Now, we rarely use this statistic by itself — but that difference, $X_i - \bar{X}$, is integral to much of the statistical tests we will use. Look for this difference in other equations!

In R, we can get the mean deviation with the `mad()` function. At the R prompt, then

```
mad(x)
[1] 2.9652
```

Population variance and population standard deviation

The population variance is appropriate for describing variability about the middle for a **census**. Again, a census implies every member of the population was measured. The equation of the population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{N}$$

The population standard deviation also describes the variability about the middle, but has the advantage of being in the same units as the quantity (i.e., no "squared" term). The equation of the population standard deviation is

$$\sigma = \sqrt{\sigma^2}$$

Sample variance and sample standard deviation

The above statements about the population variance and standard deviation hold for the sample statistics. The equation of the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})}{n-1}$$

and the equation of the sample standard deviation is

$$s = \sqrt{s^2}$$

Of course, instead of taking the square-root of the sample variance, the sample standard deviation could be calculated directly.

$$s = \sqrt{\frac{X_i - \bar{X}}{n-1}}$$

Note:

See the difference between calculation of the population parameter and the sample statistic estimates? The difference between the formulas for population and sample variances — We divide by `n - 1` instead of `N`. This is **Bessel's correction** and we will take a few moments here and in class to show you why this correction makes sense. Bessel's correction to the sample variance illustrates a basic issue in statistics: when estimating something, we want the estimator (i.e., the equation), to be an unbiased value for the population parameter it is intended to estimate.

Statistical bias is an important concept in its own right; bias is a problem because it refers to a situation in which an estimator consistently returns a value different from the population parameter for which it is intended to estimate. Thus, the sample mean is said to be an **unbiased estimator** of the population mean. Here, bias means that the formula will give you a good estimate of the value. This turns out not to be the case for the sample variance if you divide by `n` instead of `n - 1`. Now, for very large values of `n`, this is not much of an issue, but it shows up when `n` is small. In R you get what you ask for — if you ask for the sample standard deviation, the software will return the correct value; calculators, go to watch out for this — not all of them are good at communicating which standard deviation they are calculating, the population or the sample standard deviation.

Winsorized variances

Like the trimmed mean and winsorized mean (Chapter 3.2), we may need to construct a robust estimate of variability less sensitive to extreme **outliers**. Winsorized refers to procedures to replace extreme values in the sample with a smaller value. As noted in Chapter 3.2, we chose the level ahead of time, e.g., 90%. Winorized values then The **winsorized variance** is just the sample variance of the winsorized values. In R, we use `winvar()` from the WRS2 package.

Making use of the sample standard deviation

Given an estimate of the mean and an estimate of the standard deviation, one can quickly determine the kinds of observations made and how frequently they are expected to be found in a sample from a population (assuming a particular population distribution). For example, it is common in journal articles for authors to provide a table of summary statistics like the mean and standard deviation to describe characteristics of the study population (aka the reference population), or samples of subjects drawn from the study population (aka the sample population). The CDC provides reports of attributes of for a sample of adults (more than forty thousand) from the USA population (Fryar et al 2018). Table 3.3.1 shows a sample of results for height, weight, and waist circumference for men aged 20 – 39 years.

Table 3.3.1. Summary statistics mean (± standard deviation) of height, weight, and waist circumference of 20-39 year old men, USA.

Years	Height, inches	Weight, pounds	Waist Circumference, inches
1999 – 2000	69.4 (0.1)	185.8 (2.0)	37.1 (0.3)
2007 – 2008	69.4 (0.2)	189.9 (2.1)	37.6 (0.3)
2015 – 2016	69.3 (0.1)	196.9 (3.1)	38.7 (0.4)

We introduced the normal deviate as a way to normalize scores, and which we we will use extensively in our discussion of the normal distribution in Chapter 6.7, as a way to standardize a sample distribution, assuming a normal curve.

$$Z = \frac{X_i - \mu}{\sigma}$$

For data that are normally distributed, the standard deviation can be used to tell us a lot about the variability of the data:

62.26% of the data will lie between $\pm 1 \cdot \sigma$ of $\bar{X}$

95.46% of the data will lie between $\pm 2 \cdot \sigma$ of $\bar{X}$

99.0% of the data will lie between $\pm 3 \cdot \sigma$ of $\bar{X}$

This is known as the **empirical rule**, where 68% of the observations will be within one standard deviation, 95% of observations will be within two standard deviations, and 99% of observations will be within three standard deviations.

For example, men aged 20 years in the USA are on average μ = 5 feet 11 inches tall, with a standard deviation of σ = 3 inches. Consider a sample of 1000 men from this population. Assuming a normal distribution, we predict that 623 (62.26%) will be between 5 ft. 8 in. and 6 ft. 2 in., or $\pm 1 \cdot \sigma$.

Where did the 5 ft. 8 in. and the 6 ft. 2 in. come from? We add or subtract multiples of standard deviations. Thus, 6 ft. 2 in. = 5 ft. 11 in. $+1 \cdot \sigma$ (replace σ with 3 in.) and 5 ft. 8 in. = 5 ft. 11 in. $-1 \cdot \sigma$ (again, replace σ with 3 in.).

We can generalize to any distribution. **Chebyshev's inequality** (or theorem) guarantees that no more than a particular fraction $1/k^2$ of observations can be a specified k standard deviations distance away from the mean (k needs to be greater than 1). Thus, for $k = 2$ standard deviations, we expect a minimum of 75% of values $\left(1 - 1/2^2\right)$ within two standard deviations away from the mean, or for $k = 3$, then 89% of values $\left(1 - 1/3^2\right)$ will be within three standard deviations from the mean.

Hopefully you are now getting a sense how this knowledge allows you to plan an experiment.

For example, for a sample of 1000 observations of height of men, how many do we expect to be greater than 6 ft. 7 in. tall? Apply the empirical rule and do a little math — 79 inches (6 ft. 7 in) minus 71 inches (the mean) is equal to 8. Divide 8 by 3 (our value of σ) and you'll get 2.66667; so, 6 ft. 7 in. tall is about 2.67 standard deviations greater than the mean. From Chebyshev's inequality we have $1/2.67^2 = 0.140$, or 14% of observations less than or greater than the mean $(\pm\mu)$. Our question asks only about expected number of observations greater than $+2.67 \cdot \sigma$; divide 14% in half — we therefore expect about 70 individuals out of 1000 men

sampled to be 6 ft. 7 in. or taller. Note that we made no assumption about the shape of the distribution. If we assume the sample of observations comes from the normal distribution, then we can apply the Z score: about 4 individuals out of 1000 are expected to be $+2.67 \cdot \sigma$ from the mean (see R code in Note box). We extend the Z-score work in Chapter 6.7.

Note:

R code:

```
round(1000*(pnorm(c(79), mean=71, sd=3,lower.tail=FALSE)),0)
```

or alternatively

```
round(1000*(pnorm(c(2.67), mean=0, sd=1,lower.tail=FALSE)),0)
```

R output

```
[1] 4
```

R Commander command

Rcmdr: **Distributions** → **Continuous distributions** → **Normal distribution** → **Normal probabilities …**, select Upper tail.

For another example, this time we will use a data set available in R: the diabetic retinopathy data set in the survival package. It contains ages of the 394 subjects ranged from 1 to 58. Mean was 20.8 ± 14.81 years. How many out of 100 subjects do we expect to be greater than 50 years old? With Chebyshev's inequality we have $1/1.97^2 = 0.257$, or 25.7% of observations less than or greater than the mean, so about 13. If we assume normality, then the Z score (Upper tail) is 28.5% and we expect 28 subjects older than 50. Checking the data set, only eight subjects were older than 50. Our estimate by Chebyshev's inequality was closer to the truth. Why? Take a look at the histogram of the ages.

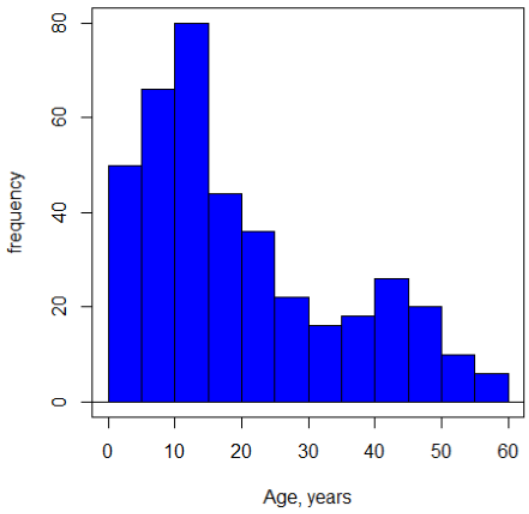


Figure 3.3.3: Histogram of subject ages from diabetic retinopathy data set in the R survival package.

Doesn't look like a normal distribution, does it?.

Z-score or Chebyshev's inequality, which to use? Chebyshev's inequality is more general — it can be used whether the variable is discrete or continuous, and without assumption about the distribution. In contrast, the Z score assumes more is known about the variable: random, continuous, drawn from a normally distributed population. Thus, as long as these assumptions hold, the Z score approach will give a better answer. This makes intuitive sense — if we know more, our predictions should be better.

In summary from the above points, and perhaps more formally for our purposes, the standard deviation is a good statistic to describe variability of observations on subjects, it's integral to the concept of precision of an estimate and is part of any equation

for calculating confidence intervals (CI). For any estimated statistic, a good rule of thumb is to always include a confidence interval calculation. We introduced these intervals in our discussion of risk analysis (approximate CI of NNT), and we will return to confidence intervals more formally when we introduce *t*-tests.

When you hear people talk about "**margin of error**" in a survey, typically they are referring to the standard deviation — to be more precise, they are referring to a calculation that includes the standard deviation, the standard error and an accounting for confidence in the estimate (see also Chapter 3.5 – Statistics of error).

Corrected Sums of Squares

Now, take a closer look at the sample variance formula. We see a *deviate*:

$$X_i - \bar{X}$$

The variance is the average of the squared deviations from the mean. Because of the "squared" part, this statistic will always be positive and greater (or typically, equal to) zero. The variance has squared units (e.g., if the units were grams, then the variance units are grams^2). The sample standard deviation has the same units as the quantity (i.e., no "squared" term). The numerator is called a **sums of squares**, and will be abbreviated as SS. Much like the deviate, it will show up frequently as we move forward in statistics (e.g., it's key to understanding ANOVA).

Other standard deviations of means

Just as we discussed for the arithmetic average, there will be corresponding standard deviations for the other kinds of means. With the geometric mean, one would calculate the **standard deviation of the geometric mean**, s_{gm}, as

$$s_{gm} = \pm exp\sqrt{\frac{\sum_{i=1}^{n}\left(\ln\frac{X_i}{\bar{X}_{gm}}\right)^2}{n-1}}$$

where exp is the exponential function, ln refers to natural logarithm, and $\bar{X}_{gm}$ refers to the sample geometric mean.

For the sample harmonic mean, it turns out there isn't a straightforward formula, only an approximation (which I will spare you — it involves use of expectations and moments).

Base R, and therefore Rcmdr, doesn't have built in functions for these, although you could download and install some R packages which do (e.g., package `NCStats`, and the function is `geosd()`).

If we run into data types appropriate for the geometric or harmonic means and standard deviations I will point these out; for now, I present these for completeness only.

Coefficient of variation (CV)

An unfortunate property of the standard deviation is that it is related (= "correlated") to the mean. Thus, if the mean of a sample of 100 individuals is 5 and variability is about 20%, then the standard deviation is about 1; compare this to a sample of 100 individuals with mean = 25 and 20% variability, where the standard deviation is about 5. For means ranging from 1 to 100, here's a plot to show you what I am talking about.

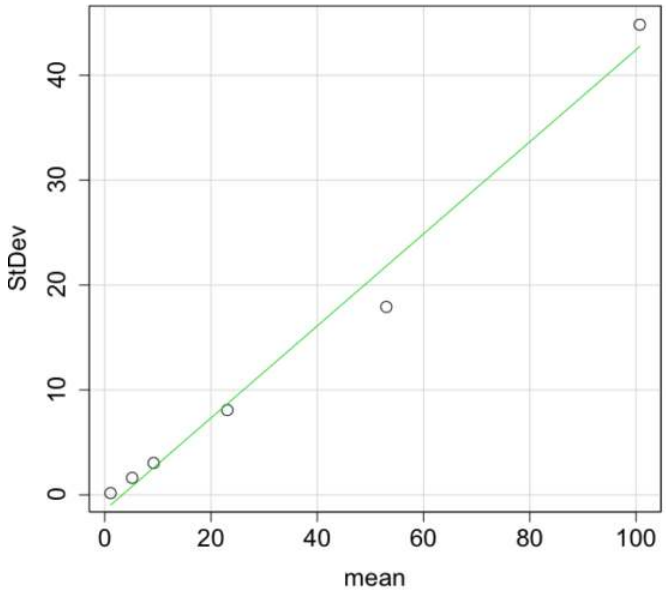


Figure 3.3.4: Scatter plot of the standard deviation (StDev) by the mean. Data sets were simulated.

The data presented in the Fig. 3.3.4 graph were simulated. Is this bias, a correlation between mean and standard deviation, something you'd see in "real" data? Here's a plot of height in inches at withers for dog breeds (Fig. 3.3.5). A line is drawn (ordinary linear regression, see Chapter 12) and as you can see, as the mean increases, the variability as indicated by the standard deviation also increases.

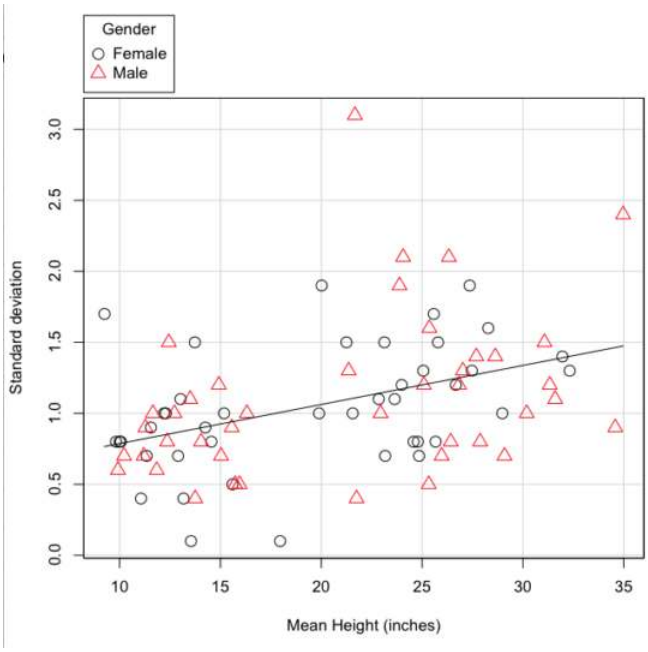


Figure 3.3.5: Plot of the standard deviation by the mean for heights of different breeds of dogs.

So, to compare variability estimates among groups when the means also vary, we need a new statistic, the coefficient of variation, which is abbreviated as CV. Many statistics textbooks not aimed at biologists do not provide this estimator (e.g., check your statistics book).

$$CV = \frac{s}{\bar{X}} \cdot 100$$

This is simply the standard deviation of the sample divided by the sample mean, then multiplied by 100 to give a percent. The CV is useful when comparing the distributions of different types of data with different means.Many times, the standard deviation of a distribution will be associated with the mean of the data.

Example. The standard deviation in height of elephants will be larger on the centimeter scale than the standard deviation of the height of mice. However, the amount of variability RELATIVE to the mean may be similar between elephants and mice. The CV might indicate that the relative variability of the two organisms is the same.

The standard deviation will also be influenced by the scale of measurement. If you measure on the millimeter scale versus the meter scale, the magnitude of the SD will change. However, the CV will be the same!

By dividing the standard deviation by the mean, you remove the measurement scale dependence of the standard deviation and generally, you also remove the relationship of the standard deviation with the mean. Therefore, CV is useful when comparing the variability of the data from distributions with different means.

One disadvantage is that the CV is only useful for ratio scale data (i.e., those with a true zero).

The coefficient of variation is also one of the statistics useful for describing the precision of a measurement. See Chapter 3.4: Estimating parameters.

Standard error of the mean (SEM)

All estimates should be accompanied by a statistic that describes the accuracy of the measure. Combined with the confidence interval (Chapter 3.4), one such statistic is called the standard error of the mean, or SEM for short. For the error associated with calculation of our sample mean, it is defined as the sample standard deviation divided by the square root of n, the sample size.

$$sem \rightarrow s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

The concept of error for estimates is a crucial concept. All estimates are made with error and for many statistics, a calculation is available to estimate the error (see Chapter 3.4).

Although related to each other, the concepts of sample standard deviation and sample standard error have distinct interpretations in statistics. The standard deviation quantifies the amount of variation of observations from the mean, while the standard error quantifies the difference between the sample mean and the population mean. The standard error will always be smaller than the standard deviation and is best left for reporting accuracy of a measure and statistical inference rather than description.

Questions

1. For a sample data set like `y = c(1,1,3,6)` , you should now be able to calculate, by hand, the
   - range
   - mean
   - median
   - mode
   - standard deviation
2. If the difference between Q1 and Q3 is called the interquartile range, what do we call Q2?
3. For our example data set, `x <- c(4,4,4,4,5,6,6,6,7,7,8,8,8,8,8)` calculate
   • IQR
   • sample standard deviation, *s*
   • coefficient of variation
4. Use the `sample()` command in R to draw samples of size 4, 8, and 12 from your example data set stored in `x` . Repeat the calculations from question 3. For example `x4 <- sample(x,4)` will randomly select four observations from *x*, and will store it in the object *x4*, like so (your numbers probably will differ!) `x4 <- sample(x,4); x4 [1] 8 6 8 6`
5. Repeat the exercise in question 4 again using different samples of 4, 8, and 12. For example, when I repeat sample(`x` ,4) a second time I get `sample(x,4) [1] 8 4 8 6`
6. For Table 1, determine how many multiples of the standard deviation for observations greater than 95-percentile (e.g., determine the observation value for a person who is in the 95-percentile for Height in the different decades, etc.
7. Calculate the sample range, IQR, sample standard deviation, and coefficient of variation for the following data sets
   - Basal 5 hour fasting plasma glucose-to-insulin ratio of four inbred strains of mice,
     `x <- c(44, 100, 105, 107) #(data from Berglund et al 2008)`

- Height in inches of mothers,
  ```
  mom <- c(67, 66.5, 64, 58.5, 68, 66.5) #(data from GaltonFamilies in R package HistData)
  ```
  and fathers,
  ```
  dad <- c(78.5, 75.5, 75, 75, 74, 74) #(data from GaltonFamilies in R package HistData)
  ```
- Carbon dioxide (CO_2) readings from Mauna Loa for the month of December for years <-c (1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010, 2015, 2020)
  ```
  years <-c (1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010, 2015, 2020) #obviously, do not calculate statistics on years; you can use to make a plot
  co2 <- c(316.19, 319.42, 325.13, 330.62, 338.29, 346.12, 354.41, 360.82, 396.83, 380.31, 389.99, 402.06, 414.26) #(data from Dr. Pieter Tans, NOAA/GML (gml.noaa.gov/ccgg/trends/) and Dr. Ralph Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/))
  ```
- Body mass of *Rhinella marina* (formerly *Bufo marinus*)
  ```
  bufo <- c(71.3, 71.4, 74.1, 85.4, 85.4, 86.6, 97.4, 99.6, 107, 115.7, 135.7, 156.2)
  ```

3.4: Estimating parameters

Introduction

What you will find on this page is definitions for **constants**, **variables**, and **parameters** as used in data analytics. **Statistical bias** and the concept of **unbiased estimators** are discussed. The different types of error are introduced along with definitions of **accuracy** and **precision**. Statistics used to quantify error are presented in the next section, Chapter 3.5.

Estimating parameters

Parameters belong to populations; they are fixed but unknown. Parameters are characteristics of populations: the typical average height and weight of five-year-old-children; average and range of gene expression of TP53 in epithelial lung cells of 50-year-old non-smoking humans; etc. Because we generally do not have at our disposal surveys of entire populations for these characteristics, we sample and calculate descriptive statistics for a subset of individuals from populations: these statistics, the mean height, range of height, etc., are **variables**. We expect average height, range of height, etc., to change, to vary, from sample to sample. A **constant**, as the word implies, is not variable and remains unchanged.

We estimate parameters (means, variances) from samples of **observations** from a **population**. Intuitively then, our estimates are only as good as how representative of the population the sample is. Statistics allows us to be more precise: we can define "how good" our estimates are by asking about, and quantifying, the **accuracy** and **precision** of our estimates.

Not withstanding the notion of "personalized medicine," our goal in science is to understand cause and effect among sets of **observations** on **samples** from a **population**. We ask, what is the link between lung cancer and smoking? We know that smoking tobacco cigarettes increases risk of cancer, but not everyone who smokes will get cancer (Pesch et al., 2012). Tumorigenic risk is in part mediated by heredity (cf. Trifiletti et al 2017). That's one way we run into problems in statistics: the biological phenomenon is complicated in ways we are not yet aware of, and thus samples drawn from populations are heterogeneous. Put another way, we think there is one population, but there may be many distinct populations. By chance, repeated samples drawn from the population include individuals with different risk associated with heredity.

As an aside, this is precisely why the concept of personalized medicine is important. Medical researchers have learned that some people with breast cancer respond to treatment with trastuzumab, a monoclonal antibody, but others do not (review in Valabrega et al 2007). Responders have more copies of a gene, copy number variation, for an EGF-like receptor called HER2. In contrast, non-responders to trastuzumab have fewer copies of this receptor (and are said to be HER2-negative in an antibody test for HER2). Thus, we speak about people with breast cancer as if the disease is the same, and from a statistical point of view, we would assume that individuals with breast cancer are of the same population. But they are not — and so the treatment fails for some, but works for others. In this example, we have a **mechanism** or **cause** to explain why some do not respond; their breast cancer is not associated with increased copy numbers of HER2. And so in studies with breast cancer, statisticians may account for differences in HER2 status. Note: Fewer than 30% of breast cancer patients may have over-expression of HER2, (Bilous et al. 2003).

Clearly, HER2 status would be used by statisticians when drawing samples from the breast cancer patient population, and we would not mistake samples. But, in many other situations we are not aware of any differences and so we assume our samples come from the same population. Note that even with imperfect knowledge about samples, experimental design is intended to help mediate heterogenous samples. For example, this is why treatment controls need to be used, or case controls are included in studies.

Random assignment to groups also is an attempt to control for unknowns: if random, then all treatment groups will likely include representatives of the numerous co-factors that contribute to risk.

Statisticians have an additional, technical burden: there are often a variety of ways (algorithms) to describe or make inferences about samples, and they are not equally capable of giving us "truth." Estimates of a parameter are not going to be exactly the true value of the parameter! This is the problem of identifying unbiased ways to estimate parameters. In statistics, bias quantifies whether an algorithm to calculate a particular statistic (e.g., the mean or variance) is consistently too low or too high.

Random sampling

Random sampling from the population is likely to be our best procedure for obtaining representative samples, but it is not foolproof (we'll return to situations where random sampling fails to provide adequate samples of populations in Chapter 5.5: Importance of randomization in experimental design). However, a second concern is how to calculate the mean, how to calculate the variance. In the section on Descriptive Statistics (Chapter 3.2), we presented how to calculate the arithmetic mean — the simple

average — but also introduced you to other calculations for the middle (e.g., median, geometric mean, harmonic mean). How to know which "middle" is correct?

Statisticians use the concept of **bias** — the simple arithmetic mean is an unbiased estimator of the population parameter, but a correction needs to be made to the calculation of sample variance to remove bias. Bias implies that the estimator systematically misses the target in some way. Good or "best" parameter estimates are, on average, going to be close to the actual parameter if we collected many samples of the population (this will depend on the sample size and the variability of the data). Earlier we introduced Bessel's correction to the sample variance, divide the sum of squares by $n-1$ and not N as we did for the population variance, so that it is an **unbiased estimator** of the sample variance. Thus, statisticians have worked out how well their statistical equations work as estimators; concepts of **expected values** or the expectation are used to evaluate how well an estimator works. In statistics, the expected value of a statistic is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values.

We need to add a bit more to our discussion about measurement and estimation.

Types of error

To **measure** is to assign a number to something, a variable. Measurements can have multiple levels, but are of the four data types: **nominal**, **ordinal**, **interval scale**, and **ratio scale**. Nominal and ordinal types are categorical or qualitative; interval and ratio are continuous or quantitative. Errors in measurement represent differences between what is observed and recorded compared with the **true value**, i.e., the actual population value. With respect to measurement we distinguish between random errors and systematic errors. Random errors occur by chance, and are thus expected to vary from one measurement instance from another. Random error may lead to measures larger or smaller than the true value. Systematic errors are of a kind that can be attributed to failures of an experimental design or instrument. **Accuracy** is reflected in the question: how close to the true value is our measure? Accuracy is distinct from **precision**, the concept of how clustered together repeated values are for the same measurement. All measurement has associated **error**, which may be divided into two kinds: **random error** and **systematic error**.

We tend to think of error in terms of mistakes, mistakes by us or as failure of the measurement process. However, in biology research, that is too restrictive of a meaning for error. Error in biology ranges from mistakes in data collection to real differences among individuals for a characteristic. The latter source, error among individuals, is of course, not a mistake, but rather, it's the "very spice of life" (Cowper 1845). We'll leave the study of individual differences, biological error, for later. This section is concerned with error in the sense of mistakes.

Random error includes things like chance error in an instrument leading to different repeat measures of the same thing and to the reality that individual differences exist for most biological traits. We minimize the effects of this kind of error by randomizing: we randomly select samples from populations; we randomly assign samples to treatment groups. Random error can make it hard to differentiate treatment effects. Random error decreases precision, the repeatability of measures. At worse, random error is conservative — it tends to mean we miss group differences, we conclude that the treatment (e.g., aspirin analgesics) has no effect on the condition (e.g., migraines). This kind of error is referred to as a **Type II error** (Chapter 8).

The experimental design remedy for random error is to increase sample size, a key conclusion drawn from power analysis on experiments, discussed in Chapter 11. The other type of error, **systematic error**, a type of bias, is more in line with the idea of errors being synonymous with mistakes. Uncalibrated instruments yield incorrect measures. And these kinds of errors lead us to make errors that can be more problematic. An example? Back in the early 1990s when I started research on whole animal metabolic rates (e.g., Dohm et al 1994, Beck et al 1995), we routinely set baseline carbon dioxide, CO_2, levels to 0.035% of the volume of dry air (350 ppm, parts per million), which reflected ambient levels of CO_2 at the time (see Figure 4 in Chapter 4.6).

```
#NOAA monthly data from Mauna Loa Observatory
co2.1994 <- c(358.22, 358.98, 359.91, 361.32, 361.68, 360.80, 359.39, 357.42, 355.63,
```

You are probably aware, today, background CO_2 have increased considerably even since the 1980s. Data for 2018 are provided below

```
co2.2018 <- c(407.96, 408.32, 409.41, 410.24, 411.24, 410.79, 408.71, 406.99, 405.51,
```

Thus, if I naively compared rates of CO_2 produced by an animal at rest, $\dot{V}_{CO_2}$, measured today against the data I gathered back in the 1990s without account for the change in background CO_2 in my analysis, I will have committed a systematic error: values of $\dot{V}_{CO_2}$ on animals today would be systematically higher than values for 1988.

Examples of error

One under-reported cost of next-generation sequencing technologies is that base calls, the process of assigning a nucleotide base to chromatogram peaks, has more errors than traditional Sanger-based sequencing methods (Fox et al 2014). In "next-generation-sequencing" (NGS) methods, individual DNA fragments are assigned sequences, whereas Sanger methods took the average sequence of a collection of DNA fragments; thus, in principle, NGS methods should be able to characterize variation of mixtures of sequences in ways not available to traditional sequencing approaches. However, artifacts introduced in sample preparation and PCR amplification lead to base calling errors (Fox et al 2014).

Gene expression levels will differ among tissue types, thus mixed samples from different tissues will misrepresent gene expression levels.

Measurement of energy expenditure, oxygen uptake, and carbon dioxide release by an organism have long been sources of study in ecology and other disciplines. A classic measure is called basal metabolic rate, measured as the rate of oxygen consumption of an endothermic animal, post-absorptive (i.e., not digesting food), at rest but not sleeping, while the animal is contained within a thermal-neutral environment (Blaxter 1989).

Accuracy and precision

Two properties of measurement are the accuracy of the measure and its precision. **Accuracy** is defined as the closeness of a measured value to its true value. **Precision** refers to the closeness of a second measure to the first, to the closeness of a third measure to the first and second, and so on. Precision refers to repeatability of measurement. We suggested use of the coefficient of variation, defined with examples in Chapter 3.2, as a way to quantify precision.

Consider the accuracy and precision of three volumetric pipettors: p1000, which has a nominal range between 100 and 1000 µL (microliters); p200, which has a range between 20 and 200 µL; p100, which has a range between 10 and 100 µL. Which of these three pipettors do you think would have the best accuracy and precision for dispensing 100 µL? We can test pipettors by measuring the mass of distilled water dispensed by the pipettor on an analytical balance. For 100 µL of distilled water at standard temperature and pressure conditions, the mass of the water would be 0.100 grams. The results are shown in the table, and a dot plot is shown in the figure to help us see the numbers.

Table 3.4.1. Mass (grams) of 100 µL of distilled water dispensed by three volumetric pipettors*.

	p1000	**p200**	**p100**
	0.113	0.100	0.101
	0.114	0.100	0.100
	0.113	0.100	0.100
	0.115	0.099	0.101
	0.113	0.100	0.101
	0.112	0.100	0.100
	0.113	0.100	0.100
	0.111	0.100	0.100
	0.114	0.101	0.101
	0.112	0.100	0.100
mean:	0.113	0.100	0.1004
standard deviation:	0.0012	0.0005	0.0005

*Temperature 21.5 °C, barometric pressure 76.28 cm mercury (elevation 52 meters). Data presented in this table were not corrected to standard temperature or pressure.

Here is the code for R; assuming that you have started R commander, then copy and paste each line into the R Commander script window; if not, enter the script one line at a time in the R console. Recall that the hashtag, #, is used to add comment lines.

A reminder: Don't include the last two rows from Table 3.4.1 in your data set; these contain descriptive statistics and are not your data.

```
#create the variables
p1000 <- c(0.113, 0.114, 0.113, 0.115, 0.113, 0.112, 0.113, 0.111, 0.114, 0.112)
p100 <- c(0.101, 0.1, 0.1, 0.101, 0.101, 0.1, 0.1, 0.1, 0.101, 0.1)
p200 <- c(0.1, 0.1, 0.1, 0.099, 0.1, 0.1, 0.1, 0.1, 0.101, 0.1)
```

```
#create a data frame
pipet <- data.frame(p100,p200,p1000)
attach(pipet)
```

```
#stack the data
stackPipet <- stack(pipet[, c("p100","p200","p1000")])
#Add variable names
names(stackPipet) <- c("mass", "pipet")
#create the dot plot
with(stackPipet, RcmdrMisc::Dotplot(mass, by=pipet, bin=FALSE))
```

From Table 3.4.1 we see that the means for the p100 and p200 were both close to the target mass of 0.1 g. The mean for the p1000, however, was higher than the target mass of 0.1 g. A dot plot is a good way to display measurements (Fig. 3.4.1).

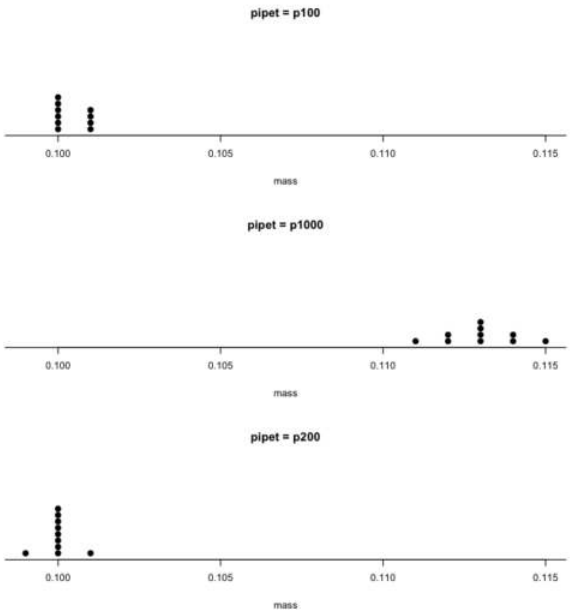


Figure 3.4.1: Dot plot of pipet results.

Note:

`Dotplot()` is part of the RcmdrMisc package; if you are using the Rcmdr script window (please do!), then the functions in RcmdrMisc are already available and you wouldn't need to use `RcmdrMisc::` (package name plus double-colon operator) to call the `Dotplot()` function (more properly, referred to as **namespaces**). Thus, `with(stackPipet, Dotplot(mass, by=pipet, bin=FALSE))` would work perfectly well within the Rcmdr script window.

From the dot plot we can quickly see that the p1000 was both inaccurate (the data fall well above the true value), and lacked precision (the values were spread about the mean value). The other two pipettors showed accuracy and looked to be similar in precision, although only two of ten values for the p200 were off target compared to four of ten values for the p100. Thus, we would conclude that the p200 was best at dispensing 100 microliters of water.

In the next chapter we apply statistics to estimate our confidence in this conclusion. In contrast to measurement, **estimation** implies calculation of a value. In statistics, estimates may be a **point**, e.g., a value of a collection of data, or an **interval**, e.g., a confidence interval.

Summary

This is your first introduction into the concept of experimental design, as defined by statisticians! One of the key tasks for a statistical analyst is to have an appreciation for measurement accuracy and precision as established in the experiment. Precise and accurate measurement levels determine how well questions about the experiment can be answered. At one extreme, if a measure is imprecise, but accurate, then it will be challenging to quantify differences between a control group and a treatment group. At the other extreme, if the measure is precise, but inaccurate, the danger would be differences between the treatment and control group may be more likely, even when the groups are truly not different!

Questions

1. List the types of error in measuring mRNA expression levels on a gene for a sampling of cells from biopsy of normal tissue and a biopsy of a tumor. Assume use of NEXGEN methods for measuring gene expression. Distinguish between technical errors and biological errors.
2. If confidence interval are useful for estimating accuracy, what statistic do we call to quantify precision?

3.5: Statistics of error

Introduction

In this section, following the discussion about error in statistics, you'll find a justification for use of confidence intervals, how to calculate confidence intervals, both as an approximation and with an example of exact calculation, use of confidence interval to quantify accuracy, and conclude with a brief discussion of rounding and significant figures.

Statistics of error

An **error** in statistics means there was a difference between the measured value and the actual value for an object. Classical statistical approach developed a large body of calculated statistics, e.g., standard error of the mean, which allows the user to quantify how large the error of measurement is given assumptions about the distribution of the errors. Thus, classical statistics requires user to make assumptions about the error distribution, the subject of our Chapter 6. A critical issue to understand is that these methods assume large sample sizes are available; they are called **asymptotic statistics** or **large-sample statistics**; the properties of the statistical estimates are evaluated as sample size approaches infinity.

Jackknife sampling and **bootstrap sampling** are permutation approaches to working with data when the **Central Limit Theorem** — as sample size increases, the distribution of sample means will tend to a normal distribution (see Chapter 6.7) — is unlikely to apply or, rather, we don't wish to make that assumption (Chapter 19). The jackknife is a sampling method involving repeatedly sampling from the original data set, but each time leaving one value out. The estimator, for example, the sample mean, is calculated for each sample. The repeated estimates from the jackknife approach yield many estimates which, collected, are used to calculate the sample variance. Jackknife estimators tend to be less biased than those from classical asymptotic statistics.

Bootstrapping, and not jackknife resampling, may now be the preferred permutation approach (e.g., Google Scholar search "bootstrap statistics" 36K hits; "jackknife statistics" 17K hits), but which method is best depends on qualities of the data set. Bootstrapping involves large numbers of permutations of the original data, which, in short, means we repeatedly take many samples of our data and recalculate our statistics on these sets of sampled data. We obtain statistical significance by comparing our result from the original data against how often results from our permutations on the resampled data sets exceed the originally observed results. By permutation methods, the goal is to avoid the assumptions made by large-sample statistical **inference**, i.e., reaching conclusions about the population based on samples from the population. Since its introduction, "bootstrapping" has been shown to be superior in many cases for statistics of error compared to the standard, classical approach (add citations).

There are many advocates for the permutation approaches, and, because we have computers now instead of the hand calculators our statistics ancestors used, permutation methods may be the approach you will take in your own work. However, the classical approach has its strengths — when the conditions, that is, when the assumptions of **asymptotic statistics** are met by the data, then the classical approaches tend to be less **conservative** than the permutation methods. By conservative, statisticians mean that a test performs at the level we expect it to. Thus, if the assumptions of classical statistics are met they return the correct answer more often than do the permutation tests.

Error and the observer

Individual researchers make observations, therefore, we can talk about observer variation as a kind of error measurement. For repeated measures of the same object by an individual, we would expect the individual to return the same results. To the extent repeated measures differ, this is **intraobserver error**. In contrast, measures of the same object from different individuals is **interobserver error**. For a new instrument or measurement system, one would need to establish the reliability of the measure: confronted with the same object, do researchers get the same measurement? Accounting for interobserver error applies in many fields, e.g., histopathology of putative carcinoma slides (Franc et al 2003), liver biopsies for cirrhosis (Rousselet et al 2005), blood cell counts (Bacus 1973).

Confidence in estimates

A really useful concept in statistics is the idea that you can assign how confident you are to an estimate. This is another way to speak of the accuracy of an estimate. Clearly, we have more confidence in a sample estimate for a population parameter if many observations are made. Another factor in our ability to estimate is the magnitude of observation differences. In general, the larger the differences among values from repeated trials, the less confident we will be in out estimate, unless, again, we make our estimates from a large collection of observations. These two quantities, **sample size** and **variability,** along with our level of confidence, e.g., 95%, are incorporated into a statistic called the **confidence interval**.

We will use this concept a lot throughout the course; for now, a simple but **approximate confidence interval** is to use the 2 x **SEM** rule (as long as sample size large): twice the standard error of the mean. Take your estimate of the mean, then add (upper limit) or subtract (lower limit) twice the value of the standard error of the mean (if you recall, that's the standard deviation divided by the square-root of the sample size).

$$\mu = \bar{X} \pm 2 \cdot s_{barX}$$

Example. Consider five magnetic darts thrown at a dart board (28 cm diameter, height of 1.68m from the floor) from a distance of 3.15 meters.

Figure 3.5.1: Magnetic dart board with 5 darts.

The distance in centimeters (cm) between where each of the five darts landed on the board compared to the bullseye is reported in Table 3.5.1.

Table 3.5.1. Results of five darts thrown at a target

Dart label	Distance in centimeters from center
1	7.5
2	3.0
3	1.0
4	2.7
5	7.4

Note:

Use of the coordinate plane, and including the angle measurement in addition to distance (the vector) from center, would be a better analysis. In the context of darts, determining accuracy of a thrower is an Aim-Point targeting problem and part of your calculation would be to get MOA (minute of angle). For the record, the angles (degrees) were

1. 124.4
2. -123.7
3. 96.3
4. -84.3
5. -31.5

measured using imageJ. Because there seems to be an R package for just about every data analysis scenario, unsurprisingly, there's an R package called `shotGroups` to analyze shooting data.

How precise was the dart thrower? We'll use the **coefficient of variation** as a measure of precision. Second, how accurate were the throws? Use R to calculate

```
darts = c(7.5, 3.0, 1.0, 2.7, 7.4)
#use the coefficient of variation to describe precision of the throws
coefVar = 100*(sd(darts)/mean(darts)); coefVar
[1] 68.46141
```

Confidence Interval to describe accuracy

Note that the true value would be a distance of zero — all bullseyes. We need to calculate the standard error of the mean (SEM); then, we calculate the confidence interval around the sample mean.

```
#Calculate the SEM
SEM <- sd(darts)/sqrt(length(darts)); SEM
[1] 1.322649
#now, get the lower and upper limit, subtract from the mean
confidence <- c(mean(darts)-2*SEM, mean(darts), mean(darts)+2*SEM); confidence
[1] 1.674702, 4.320000, 6.965298
```

The mean was 4.3 cm; therefore, to get the lower limit of the interval subtract 2.65 $(2 \cdot SEM = 2.645298)$ from the mean; for the upper limit add 2.65 to the mean. Thus, we report our approximate confidence interval as $(1.7, 7.0)$, and we read this as saying we are about 95% confident the population value is between these two limits. Five is a very small sample number*, so we shouldn't be surprised to learn that our approximate confidence interval would be less than adequate. In statistical terms, we would use the *t*-distribution, and not the normal distribution, to make our confidence interval in cases like this.

*Note:

As a rule, implied by **Central Limit theory** and use of **asymptotic statistical estimation**, a sample size of 30 or more is safer, but probably unrealistic for many experiments. This is sometimes called as the **rule of thirty**. (For example, a 96-well PCR array costs about $500; with $n = 30$, that's $15,000 US Dollars for one group!). So, what about this rule? This type of thinking should be avoided as "a relic of the pre-computer era," (Hesterberg, T. (2008). It's Time To Retire the" n>= 30" rule.). We can improve on asymptotic statistics by applying bootstrap principles (Chapter 19).

We made a quick calculation of the confidence interval; we can get make this calculation by hand by incorporating the *t* distribution. We need to know the **degrees of freedom**, which in this case is 4 ($n - 1$, where $n = 5$). We look up critical value of *t* at 5% (to get our 95% confidence interval), $t = 2.78$. Subtract for lower limit $t \cdot SEM$ and add for upper limit $t \cdot SEM$ to the sample mean for the 95% confidence interval. We can get help from R, by using the one-sample t-test with a test against the hypothesis that the true mean is equal to zero

```
#make an attach a data frame object
Ddarts <- data.frame(darts)
t.test(darts,mu=0)
One Sample t-test
data: darts
t = 3.2662, df = 4, p-value = 0.0309
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
 0.6477381 7.9922619
sample estimates:
mean of x
 4.32
```

`t.test` uses the function `qt()`, which provides the quantile function. To recreate the 95% CI without the additional baggage output from the `t.test`, we would simply write

```
#upper limit
mean(darts)+ qt(0.975,df=4)*sd(darts)/sqrt(5)

#lower limit
mean(darts)+ qt(0.025, df = 4, lower.tail=TRUE)*sd(darts)/sqrt(5)
```

where `sd(darts)/sqrt(5)` is the standard error of the mean.

Or, alternatively, download and take advantage of a small package called `Rmisc` (not to be confused with the RcmdrMisc package) and use the function `CI`

```
library(Rmisc)
CI(darts)
     upper     mean     lower
7.9922619 4.3200000 0.6477381
```

The advantage of using the `CI()` command from the package `Rmisc` is pretty clear; I don't have to specify the degrees of freedom or the standard error of the mean. By default, CI reports the 95% confidence interval. we can specify any interval simply by adding to the command. For example,

```
CI(darts, ci=0.90)
```

reports upper and lower limits for the 90% confidence interval.

Significant figures

And finally, we should respect **significant figures**, the number of digits which have meaning. Our data were measured to the nearest tenth of a centimeter, or two significant figures. Therefore, if we report the confidence interval as (0.6477381, 7.9922619), then we imply a **false level of precision**, unless we also report our **random sampling error of measurement.**

R has a number of ways to manage output. One option would be to set number of figures globally with the `options()` function — all values reported by R would hold for the entire session. For example, `options(digits=3)` would report all numbers to three significant figures. Instead, I prefer to use `signif()` function, which allows us to report just the values we wish and does not change reporting behavior for the entire session.

```
signif(CI(darts),2)
upper  mean lower
 8.00  4.30  0.65
```

Note:

The `options()` function allows the R user to set a number of settings for an R session. After gaining familiarity with R, the advanced user recognizes that many settings can be changed to make the session work to report in ways more convenient to the user. If curious, submit `options()` at the R prompt and available settings will be displayed.

The R function `signif()` applies rounding rules. We apply rounding rules when required to report estimates to appropriate levels of precision. Rounding procedures are used to replace a number with a simplified approximation. Wikipedia provides a comprehensive list of rounding rules. Notable rules include

- directed rounding to an integer, e.g., rounding up or down
- rounding to nearest integer, e.g., round half up if the number ends with 5
- randomly rounding to an integer, e.g., stochastic rounding.

With the exception of stochastic rounding, all rounding methods impose biases on the sets of numbers. For example, the round half up method applied for numbers above 5, round down for numbers below 5 will increase the variance of the sample. In R, use `round()` for most of your work. If you need one of the other approaches, for example, to round up, the command is `ceiling()` ; to round down we use `floor()` .

When to round?

No doubt your previous math classes have cautioned you about the problems of **rounding error** and their influence on calculation. So, as a reminder, if reporting calls for rounding, then always round after you've completed your calculations, never during the calculations themselves.

A final note about significant figures and rounding. While the recommendations about reporting statistics are easy to come by (and often very proscriptive, e.g., Table 1, Cole 2015), there are other concerns. **Meta-analysis**, which are done by collecting information from multiple studies, would benefit if more and not fewer numbers are reported, for the very same reason that we don't round during calculations.

Questions

1. Calculate the correct 90% and 99% confidence intervals for the dart data using the t-distribution
   - by hand
   - by one alternative method in R, demonstrated with examples in this page
2. How many significant figures should be used for the volumetric pipettor p1000? The p200? The p20 (data at end of this page)?
3. Another function, `round()` , can be used. Try

```
round(CI(darts),2)
```

1. - and report the results: vary the significant figures from 1 to 10 (`signif()` will take digits up to 22).
   - Note any output differences between `signif()` and `round()` ? Don't forget to take advantage of R help pages (e.g., enter `?round` at the R prompt) and see Wikipedia.
2. Compare rounding by `signif()` and `round()` for the number 0.12345. Can you tell which rounding method the two functions use?
3. Calculate the coefficient of variation (CV) for each of the three volumetric pipettors from the data at end of this page. Rank the CV from smallest to largest. Which pipettor had the smallest CV and would therefore be judged the most precise?
4. Standards distinguish between within run precision and between run precision of a measurement instrument. The data in Table 1 were all recorded within 15 minutes by one operator. What kind pf precision was measured?
5. Calculate the standard error of the means for each of the three pipettors from the data provided at end of this page.
6. Calculate the approximate confidence interval using the 2SE rule and judge which of the three pipettors is the most accurate (narrowest confidence interval)
   - Repeat, but this time apply your preferred R method for obtaining confidence intervals.
   - Compare approximate and R method confidence intervals. How well did the approximate method work?

Data sets

Pipette calibration

Table 3.5.2. Mass (grams) of 100 µL of distilled water dispensed by three volumetric pipettes.

p1000	p200	p100
0.113	0.1	0.101

p1000	p200	p100
0.114	0.1	0.1
0.113	0.1	0.1
0.115	0.099	0.101
0.113	0.1	0.101
0.112	0.1	0.1
0.113	0.1	0.1
0.111	0.1	0.1
0.114	0.101	0.101
0.112	0.1	0.1

3.6: Chapter 3 References and Suggested Reading

Bacus, J. W. (1973). The observer error in peripheral blood cell classification. *American Journal of Clinical Pathology*, *59*(2), 223-230.

Berglund, E. D., Li, C. Y., Poffenberger, G., Ayala, J. E., Fueger, P. T., Willis. S. E., Jewell, M. M., Powers, A. C., Wasserman, D. H. (2008). Glucose Metabolism In Vivo in Four Commonly Used Inbred Mouse Strains. *Diabetes* 57(7):1790-1799

Bilous, M., Ades, C., Armes, J., Bishop, J., Brown, R., Cooke, B., Cummings, M., Farshid, G., Field, A., Morey, A., McKenzie, P., Raymond, W., Robbins, P., Tan, L. (2003). Predicting the HER2 status of breast cancer from basic histopathology data: an analysis of 1500 breast cancers as part of the HER2000 International Study. *Breast* 12(2):92-98.

Blaxter, K. (1989). *Energy metabolism in animals and man*. Cambridge University Press.

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1), 2-10.

Burnett, R.W. (1975). Accurate estimation of standard deviations for quantitative methods used in clinical chemistry. *Clinical Chemistry* 21(13):1935-1938

Chambers, J. M. (2008). *Software for data analysis: programming with R* (Vol. 2). New York: Springer.

Cole, T. J. (2015). Too many digits: the presentation of numerical data. *Archives of Disease in Childhood*, *100*(7), 608-609.

Cowper, W. (1845). *The Task: And Other Poems*. Carey and Hart. Project Gutenberg January 1, 2003 [EBook #3698]

Driscoll et al (2000) An introduction to everyday statistics – 2. *Journal of Accident & Emergency Medicine* 17:274-281

Driscoll et al 2000 An introduction to everyday statistics – 2. J Accid Emerg Med2000;17:274-281

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications*, 1.

Franc, B., De La Salmonière, P., Lange, F., Hoang, C., Louvel, A., De Roquancourt, A., … & Chastang, C. (2003). Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid carcinoma. *Human pathology*, *34*(11), 1092-1100.

Fryar, C. D., Kruszan-Moran, D., Gu, Q., & Ogden, C. L. (2018). Mean body weight, weight, waist circumference, and body mass index among adults: United States, 1999–2000 through 2015–2016. *National Health Statistics Report* 122:1-16.

Koll, C.E.M., Hopff, S.M., Meurers, T. *et al.* (2022). Statistical biases due to anonymization evaluated in an open clinical dataset from COVID-19 patients. *Scientific Data* **9**, 776 . https://doi.org/10.1038/s41597-022-01669-9

Lee DK, In J, Lee S. (2015) Standard deviation and standard error of the mean. *Korean J Anesthesiol*. 68:220–223.

Pesch, B., Kendzia, B., Gustavsson, P., Jöckel, K. H., Johnen, G., Pohlabeln, H., … & Wichmann, H. E. (2012). Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case–control studies. *International journal of cancer*, 131(5), 1210-1219.

R Language Definition, version 4.1.1 (link)

Rousselet, M. C., Michalak, S., Dupré, F., Croué, A., Bedossa, P., Saint-André, J. P., & Calès, P. (2005). Sources of variability in histological scoring of chronic viral hepatitis. *Hepatology*, *41*(2), 257-264.

Trifiletti, D. M., Sturz, V. N., Showalter, T. N., & Lobo, J. M. (2017). Towards decision-making using individualized risk estimates for personalized medicine: A systematic review of genomic classifiers of solid tumors. *PloS one*, 12(5), e0176388.

Valabrega, G., Montemurro, F., Aglietta, M. (2007). Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. *Ann Oncol*. 18(6):977-984

Westgard, J. O., Carey, R. N., Wold, S. (1974). Criteria for judging precision and accuracy in method development and evaluation. *Clinical Chemistry* 20(7):825-833.

Whiteley & Ball (2002). Statistical review 1: Presenting and summarizing data. *Critical Care* 6:66-71.

CHAPTER OVERVIEW

4: How to Report Statistics

Introduction

While you are thinking about exploring data sets and descriptive statistics, please review our overview of data analysis (Chapter 2.4 and 2.5). While the scientific hypotheses come first, how experiments are designed should allow for straight-forward analysis: in other words, statistics can't rescue poorly designed experiments, nor can it reveal new insight after the fact.

Once the experiments are completed, all projects will go through a similar process.

- Description: Describe and summarize the results
- Check assumptions
- Inference: conduct tests of hypotheses
- Develop and evaluate statistical models

Clearly this is a simplification, but there's an expectation your readers will have about a project. Basic questions like how many subjects got better on the treatment? Is there an association between Body Mass Index (BMI) and the **primary outcome**? Did male and female subjects differ for response to the treatment? Undoubtedly these and related questions form the essence of the inferences, but providing graphs to show patterns may be as important to a reader as any **p-value** — a number which describes how likely it is that your data would have occurred by chance — e.g., from an Analysis of variance.

Each project is unique, but what elements must be included in a results section?

Data visualization

We describe data in three ways: graphs, tables, and in sentences. In this page we present the basics of when to choose a graph over presenting data in a table or as a series of sentences (i.e., text). In the rest of this chapter we introduce the various graphics we will encounter in the course. Chapter 4 covers eight different graphics, but is by no means an exhaustive list of kinds of graphs. Phylogenetic network graphs are presented in Chapter 20.11. Although an important element of presentation in journal articles, we don't discuss figure legends or table titles; guidelines are typically available by the journal of choice (e.g., PLOS ONE journals guidelines).

A quick note about terminology. Data visualization encompasses **charts**, **graphs** and **plots**. Of the three terms, chart is the more generic. Graphs are used to display a function or mapping between two variables; plots are kinds of graphs for a finite set of points. There is a difference among the terms, but I confess, I won't be consistent. Instead, I will refer to each type of data visualization by its descriptive name: bar chart, pie chart, scatter plot, etc. Note that technically, a scatter plot can refer to a graph, e.g., a line drawn to reflect a linear association between the two variables, whereas bar charts and pie charts would not be a graph because no function is implied.

Why display data?

Do we just to show a graph to break the monotony of page after page of text, or do we attempt to do more with graphs? After all, isn't "a picture worth a thousand words?" In many cases, yes! Graphics allow us to see patterns. **Visualization** is a key part of **exploratory data analysis**, or **data mining** in the parlance of big data. In genomics, heat maps

Graphics are complicated and expensive to do well. Text is much cheaper to publish, even in digital form. But the ability to visualize concepts, that is, to connect ideas to data through our eyes (see Wikipedia), seems to be more the cognitive goal of graphics. Lofty purpose, desirable goal. Yes, it is true that graphics can communicate concepts to the reader, but with some caution. Images distort, and default options in graphics programs are seldom acceptable for conveying messages without bias (Glazer 2011).

Here's some tips from a book on graphical display (Tufte 1983; see also Camões 2016).

Your goal is to communicate complex ideas with clarity, precision, and efficiency. Graphical displays should:

- show the data
- avoid distorting the data
- present numbers in a small space
- help the viewer's eye to compare different pieces of data
- serve a clear purpose (description, exploration, tabulation, decoration)
- be closely integrated with statistical and verbal descriptions of a data set.

We accomplish these tasks by following general principles involving scale and a commitment to avoiding bias in our presentation.

Importantly, graphs can show patterns not immediately evident in tables of numbers. See Table 4.1 for an example of a dataset, "Anscombe's quartet," (Anscombe 1973), where a picture is clearly helpful.

Table 4.1. Anscombe's data (Anscombe 1973).

X	Y1	Y2	Y3	Y4
10	8.04	9.14	7.46	6.58
8	6.95	8.14	6.77	5.76
13	7.58	8.74	12.74	7.71
9	8.81	8.77	7.11	8.84
11	8.33	9.26	7.81	8.47
14	9.96	8.10	8.84	7.04
6	7.24	6.13	6.08	5.25
4	4.26	3.10	5.39	12.50
12	10.84	9.13	8.15	5.56
7	4.82	7.26	6.42	7.91
5	5.68	4.74	5.73	6.89
Mean (±SD)	7.50 (2.032)	7.50 (2.032)	7.50 (2.032)	7.50 (2.032)

Note that the data set does not include the column summary statistics shown in the last row of the table.

The Anscombe dataset is also available in R package stats, or you can copy/paste from Table 4.1 into a spreadsheet or text file, then load the data file into R (e.g., **Rcmdr → Load data set**). Note that the data set does not include the column summary statistics shown in the last row of the table.

Before proceeding, look again at the table — See any patterns in the table?

Maybe.… Need to be careful as we humans are really good at perceiving patterns, even when no pattern exists.

Now, look just at the last row in the table, the row containing the descriptive statistics (the means and standard deviations). Any patterns?

The means and standard deviations are the same, so nothing really jumps out at you — does that mean that there are no differences among the columns, then?

But let's see what the scatter plots look like before we conclude that the columns of `Y` 's are the same (Fig. 4.1). I'll also introduce the R package `clipr` , which is useful for working with your computer's clipboard.

Definition: Term

To show clipboard history, on Windows 10/11 press Windows logo key plus V; on macOS, open Finder and select Edit → Show Clipboard.

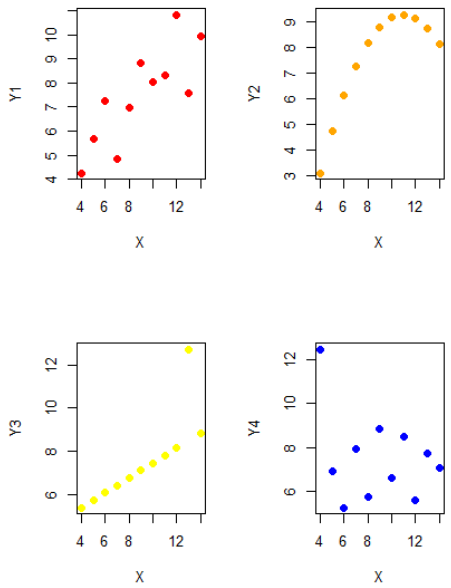


Figure 4.1: Scatter plot graphs of Anscombe's quartet (Table 4.1).

```
#R code for Figure 1.

require(clipr)
#Copy from the Table and paste into spreadsheet (exclude last row). Highlight and copy data
myTemp <- read_clip_tbl(read_clip(), header=TRUE, sep = "\t")
#Check that the data have been loaded correctly
head(myTemp)
#attach the data frame, so don't have to refer to variables as myTemp\$variable name
attach(myTemp)
#set the plot area for 4 graphs in 2X2 frame
par(mfrow=c(2,2))
plot(X, Y1, pch=19, col="red", cex=1.2)
plot(X, Y2, pch=19, col="orange",cex=1.2)
plot(X, Y3, pch=19, col="yellow",cex=1.2)
plot(X, Y4, pch=19, col="blue",cex=1.2)
```

And now we can see that the `Y` 's have different stories to tell. While the summary (descriptive) statistics are the same, the patterns of the association between `Y` values and the `X` variable are qualitatively different: `Y1` is linear, but diffuse; `Y2` is nonlinearly associated with `X` ; `Y3` , like `Y1` , is linearly related to `X` , but one data point seems to be an outlier; and for `Y4` we see a diffuse nonlinear trend and an outlier.

So, that's the big picture here. In working with data, you must look at both ways to "see" data — you need to make graphs and you also need to calculate basic descriptive statistics.

And as to the reporting of these results, sometimes tables are best (i.e., so others can try different statistical tests), but patterns can be quickly displayed with carefully designed graphs. Clearly, in this case, the graphs were very helpful to reveal trends in the data.

When to report numbers in a sentence? In a table? In a graph?

The choice depends on the message. Usually you want to make a comparison (or series of comparisons). If you are reporting one or two numbers in a comparison, a sentence is fine. "The two feral goat populations had similar mean numbers (120 vs. 125) of kids each breeding period." If you have only a few comparisons to make, the text table is useful:

Table 4.2. Data from Kipahoehoe Natural Area Reserve, SW slope of Mauna Loa.

Location	Number of kids
Outside fence:	

Location	Number of kids
Outside fence:	120
Inside fence:	
kipuka	51
Other	13
Inside fence:	
kipuka	3
Other	15

To conclude, tables are the best way to show exact numbers and tables are preferred over graphs when many comparisons need to be made. (Note: this was a real data set, but I've misplaced the citation!)

Note:

From Wikipedia, a kīpuka is a land area surrounded by recent lava flows.

Couldn't I use a pie chart for this?

Yes, but I will try to persuade you not to do so. Pie charts are used to show part-whole relationships. If there are just a few groups, and if we don't care about precise comparisons, pie charts may be effective. Sometimes, people use pie charts for very small data sets (comparing two populations, or three categories, for example). The problem with pie charts is that they require interpretation of the angles that define the wedges, so we can't be very precise about that. Bar charts (Chapter 4.1) are much better than pie charts, however.

To illustrate the problem, here's a couple of pie charts from Microsoft Excel (a similar chart can be made with LibreOffice Calc) for our goat data set; compare this graph to the table and to the bar chart below (Fig. 4.2).

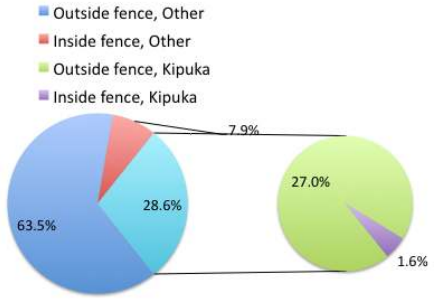


Figure 4.2: Excel pie chart of Table 2 data set.

A bar chart of the same data (Fig. 4.3):

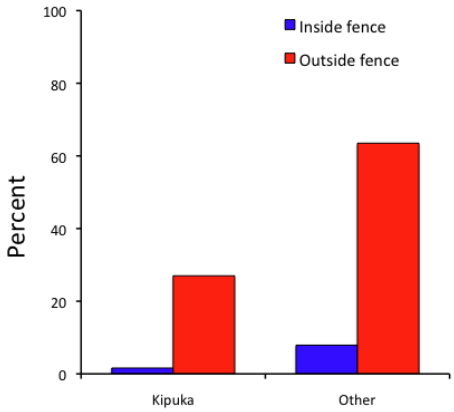


Figure 4.3: Bar chart of Table 4.2 data set.

The bar chart (Fig. 4.3) more effectively gets the message across; more goat kids were found outside the fenced area then inside the fenced in areas. We can also see that more goat kids were found in the "other" areas compared to the kipuka. The pie chart (Fig. 4.2), in my opinion, fails to communicate these simple comparisons, which are conclusions about patterns in the data that clearly would be the take-home message

from this project. Aesthetically the bar chart could be improved — a mosaic plot would work well to show the associations in the project results (See Chapter 4.4: Mosaic plots).

But we are not done with this argument on whether to use graphics or text to report results. Neither the bar chart (mosaic plot) or the pie chart really work. The reader has to interpret the graphics by extrapolating to the axes to get the numbers. While it may be boring — 1.5 million hits Google search “data tables” boring — tables can be used for comparisons and make the patterns more clear and informative to the reader. Here’s a different version of the table to emphasize the influence of fencing on the goat population.

Table 4.3. Revised Table 4.2 to emphasize comparisons between inside- and outside-the-fence-line feral goat populations on Mauna Loa.

Location	Kipuka	Other
Outside fence	51	120
Inside fence	3	15

Table 4.3 would be my choice — over a sentence and over a graph. At a glance I can see that more goat kids were found outside of the fenced area, regardless of whether it was in a kipuka or some other area on the mountain side. Table 4.3 is an improvement over Table 4.2 because it presents the comparisons in a 2 X 2 format — especially useful when we have a conditional set.

For example, it’s useful to show the breakdown of voting results in tables (numbers of votes for different candidates by voter’s party affiliation, home district, sex, economic status, etc.). Interested readers can then scan through the table to identify the comparison they are most interested in. But often, a graph is the best choice to display information. One final point: by judiciously combining words, numbers, and images, you should be able to convey even the most complex information in a clear manner! We will not spend a lot of time on these issues, but you will want to pay some attention to these points as you work on your own projects.

Some final comments about how to present data

What your graph looks like is up to you; lots of people have advice (e.g., Klass 2012). But we all know poor graphs when we see them in talks or in papers; we know them when we struggle to make sense of the take-home message. We know them when we feel like we’re missing the take-home message.

Here’s my basic take on communicating information with graphics.

- Minimize white space (for example, the scatter plots above could be improved simply by increasing the point size of the data points)
- Avoid bar charts for comparisons if you are trying to compare more than about three or four things.
- A graphic in a science report that is worth “a thousand words” probably is too complicated, too much information, and, very likely, whatever message you are trying to convey is better off in the text.

4.1: Bar (column) charts

Introduction

Bar or column charts are used to compare counts among two or more categories, i.e., an alternative to pie charts (Fig. 4.1.1).

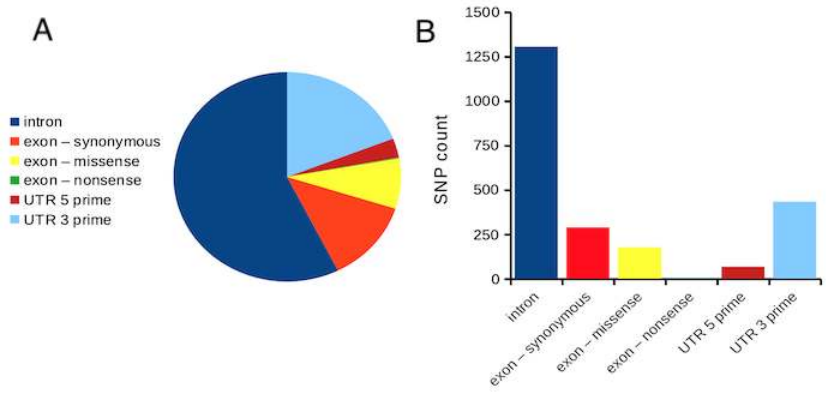


Figure 4.1.1: Single-nucleotide variants for human gene ACTB by DNA and functional element.

Although bar charts are common in the literature (Cumming et al 2007; Streit and Gehlenborg 2014), bar charts may not be a good choice for comparisons of ratio scale data (Streit and Gehlenbor 2014). Bar charts for ratio data are misleading. Parts of the range implied by the bar may never have been observed: the bars of the chart always start at zero. **Box (whisker) plots** are better for comparisons of ratio scale data and are presented in the next section of this chapter. That said, I will go ahead and present how to create bar chars for both count, generally considered acceptable, and ratio scale data, for which their use is controversial.

Purpose of the bar chart

Like all graphics, a bar chart should tell a story. The purpose of displaying data is to give your readers a quick impression of the general differences among two or more groups of the data. For counts, that's where the bar chart comes in. The bar chart is preferred over the pie chart because differences are represented by lengths of the bars in the bar chart. Differences among categories in a pie chart are reflected by angles, and it seems that humans are much better at judging lengths than angles.

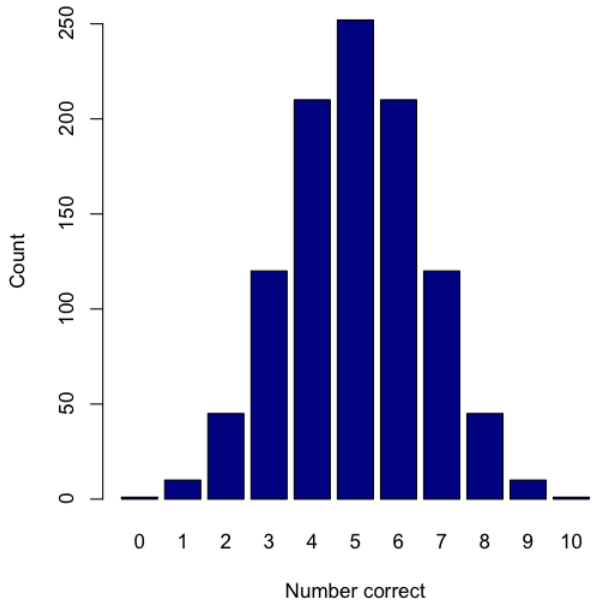


Figure 4.1.2: A simple bar chart.

```
myCombo <- seq(0,10, by=1)
myCounts <- choose(10, myCombo)    #combinations
barplot(myCounts, names.arg = myCombo, xlab = "Number correct", ylab = "Count",col =
```

A **stacked bar chart** is used to compare how different categories are further divided into subcategories shared among all the groups. For example, passengers on the Titanic at the time of its sinking can be grouped based on their passage class (first, second, or third), but if we want to compare the count of those who died or survived in each class, we can use a stacked bar chart.

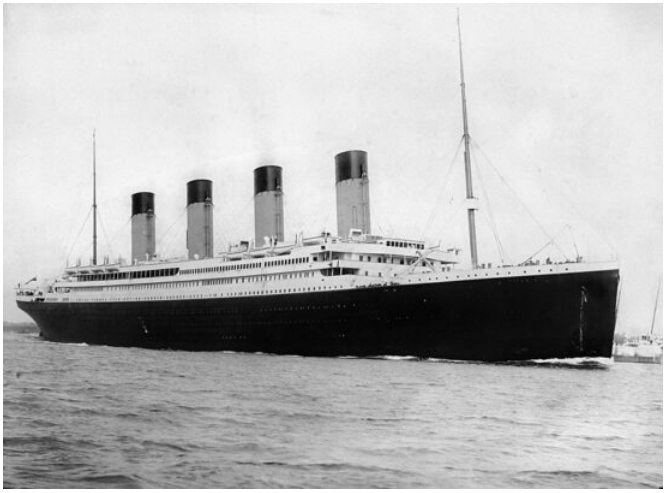


Figure 4.1.3: The luxury ship RMS *Titanic*, which sunk 15 April 1912, More than 1500 souls were lost. Public domain image, Wikipedia.

Stacked bar chart, data set `TitanicSurvival` in package `carData` .

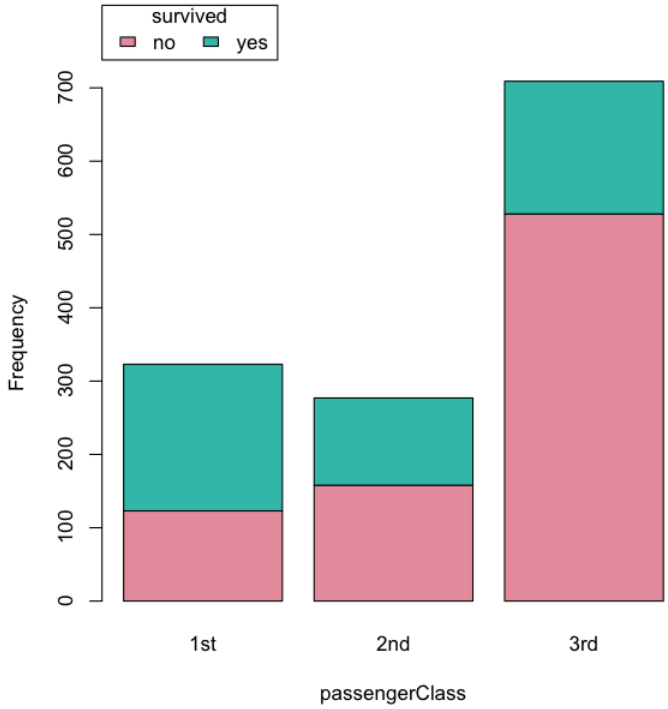


Figure 4.1.4: A stacked bar chart of survival rates on the *Titanic* by passenger class.

```
Barplot(passengerClass, by=survived, style="divided", legend.pos="above", xlab="passe
```

Bar charts with error bars

Although many data visualization specialists argue against the bar chart, their use is well established. For the familiar bar chart with ratio scale data, the X (horizontal) axis displays the categories of one variable (e.g., location, or treatment group). You plot groups to emphasize comparisons. The Y (vertical) axis then is the mean for each group.

You need **error bars**. If the mean is displayed, some measure of precision should be (must be?) displayed (Cumming et al 2007). And, as you should recall by now, your choices are **standard deviation** (Chapter 3.2), **standard error of the mean** (SEM) (Chapters 3.2, 3.5), or **confidence interval** (see Chapter 3.5). It is strongly advised that without a representation of precision, one should not interpret trends or group differences from representations of means (i.e., height of bars) alone.

The bar charts on this page are means plus or minus the standard error of the mean, ± SEM. We'll discuss which choice to make.

Examples

The Copper_rats_PMID3357063 dataset will be used for the next series of graphs (Data set). Refer to Mike's Workbook for Biostatistics Part 07 to review how to import the data.

A portion of the data set is shown below

```
head(Copper)
         Diet     Body    Heart    Liver
1 Adequate-Cu 320.1381 1.125037 10.259657
2 Adequate-Cu 329.6879 1.158982 9.843295
3 Adequate-Cu 327.9838 1.090374 9.855975
4 Adequate-Cu 334.6669 1.118183 9.942997
5 Adequate-Cu 338.3134 1.172636 9.860971
6 Adequate-Cu 345.4608 1.056183 8.885820
```

The data set consists of organ weights (heart, liver) from rats fed a diet adequate in copper, deficient in copper, and then a third group who received the adequate diet from perspective of amount of copper, but calorie restricted to match the decreased feeding rates of the rats fed the copper deficient diet. Copper is an essential trace element in our diet. The data set was simulated from descriptive statistics (means, standard deviation, number of subjects) of published data by sampling from a normal distribution. (Table 1, Ovecka et al 1988).

On we go with some graphs.

Note:

R has many options to create bar charts, and especially `ggplot2` can be used to great advantage, but there is a learning curve. One of the great things about R is that folks help each other by sharing code. For example, http://www.cookbook-r.com/Graphs/Plotting_means_and_error_bars_(ggplot2)/

But, I'm still learning about R graphics, and for bar charts with error bars, I find other packages more straight-forward. So, I'll use this moment to point out that making a good graph is more about the end product then the particular tools. I have been using other tools for years to make my graphics, so I tend to default on these options first. Graphs presented here are mostly from Veusz software program (pronounced "views"). I like it because the software allows me to edit any of the elements of the graph.

Here's a typical looking bar chart with ratio scale data: means ± SEM. Let's look at them more critically.

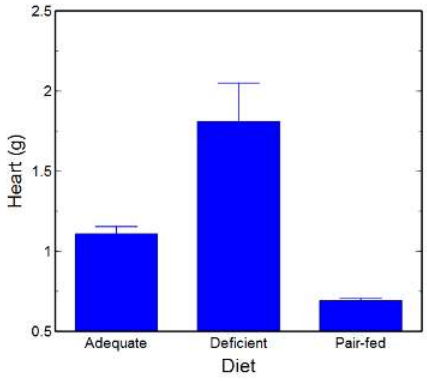


Figure 4.1.5: A bar chart with error bars (standard error of the mean).

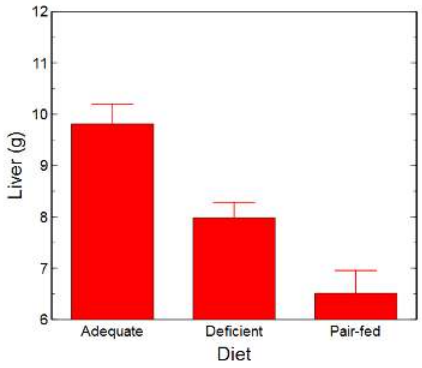


Figure 4.1.6: Another bar chart with error bars (standard errors of mean).

When making comparisons, make sure the axes have the same scale, or consider putting the graphs together.

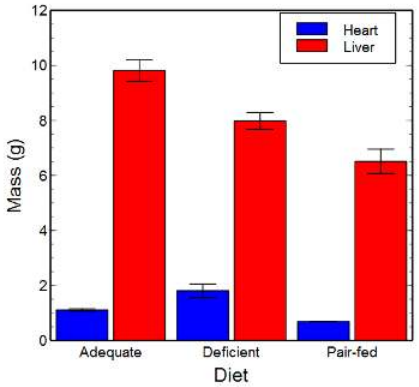


Figure 4.1.7: Bar chart that allows for a comparison among levels of a a factor (organs, liver vs. heart).

Analysis Note:

For variables that covary with body size (for example, on average, tall people generally are heavier too), a major consideration is how to present (and analyze) the data in such a way that body size is accounted for. Here, the solution was to express organ weight as the ratio of organ weight to body weight for the mice. This may or may not be a good solution, and the answer is too complicated for us now (has to do with a thing called **allometric scaling**), but I wanted to at least present the issue and show how the graphics can be improved to handle some of these concerns.

Here are the organ weights again, but taken as ratio of body mass.

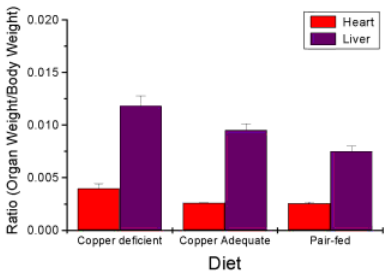


Figure 4.1.8: Same chart as in Figure 4.1.7, but in ratio form.

Note:

Let's be clear about expectations of you for statistics class. Now, R (and Rcmdr) do lots of graphics, pretty much anything you want, but it is not as friendly as it could be. For BI311 homework, the default graphics available via Rcmdr will generally be adequate for assignments. R and Rcmdr have many bar chart options, but there isn't a straightforward way to get the error bars, unless you are willing to enter some code to the command line or learn a particular package (like `gplots` or `ggplot2`).

How to make a bar chart with error bars in R

Option 1. First, let's try a work-around. Instead of an error bar option for the bar chart menu, Rcmdr provides a plot of means that allows you to plot with error bars. These are equivalent graphs, the "bar chart" and the "plot of means", though you should favor the bar chart format for publishing.

Rcmdr: Graphs → Plot of means…

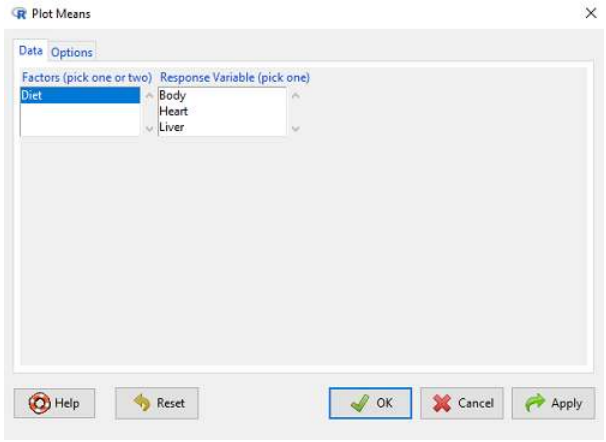


Figure 4.1.9: Rcmdr menu popup for Plot Means.

Here, Rcmdr takes the data and calculates the mean and your choice of standard errors or deviations, confidence intervals, or no error bars. The resulting graph is below.

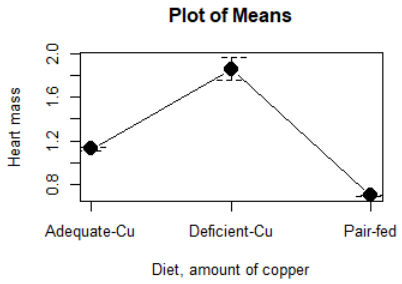


Figure 4.1.10: Plot of means, default settings.

That's an ugly graph (Fig. 4.1.10). Functional, good enough for data exploration and preliminary results, and certainly good enough for a Biostatistics homework or report. Additionally, connecting the dots here is a no-no. It implies that if we had measured categories between "adequate copper" and "deficient copper," then the points would fall on those lines. That would be a complete guess. So, why did I include the connecting lines? That was the default setting for the command, and it makes the point — think before you click. One argument for connecting points in a graph is that it makes it easier for the reader to visualize trends.

This graph (Fig. 4.1.10) is fine for exploring data, but you will want to do better for publication.

Let's make some better graphs with R

Once you are ready to go beyond the default settings available in `Rcmdr`, there is tremendous functionality in R for graphics. To access R's potential, you'll need to get into the commands a bit. I'm going to continue to try and shield you from the programming aspects of R, but from time to time you really need to see what is possible with R. Graphics is one such area. I use the package `gplots`, with 23 different graphing functions (type at R prompt `?gplots` to call up the manual pages).

`gplots` should be among the packages on your R installation; if not, then install the package and run `library(gplots)` to complete the installation. We'll try the `barchart2` function.

But first, we need to get means for each of our groups.

At the R command prompt:

```
hrtWt <- tapply(Dataset$HeartWt, list(Group=Dataset$Group), mean, na.rm=TRUE)
```

This code extracts means from our `HeartWt` variable for each Group, then stores the three (in this case, because our data set has 3 groups) in the place holder I had called `hrtWt`. To verify that the three means are there, type "hrtWt" without the quotes, then enter.

You should see

```
hrtWt Group Cu adequate Cu deficient Pair-fed
               1.200000     1.566667 0.900000
```

R functions used: `tapply`, `list`, `mean`; `na.rm` was not needed but would be used to remove all missing values (recall during our import phase we were asked how missing observations were noted in our file; the default is `NA`).

Next, I want to apply standard error bars

```
stdDEV <- tapply(Dataset$HeartWt, list(Group=Dataset$Group), sd, na.rm=TRUE)
cil <- hrtWt-(stdDEV/sqrt(3))
ciu <- hrtWt+(stdDEV/sqrt(3))
```

I used `cil` and `ciu` to designate the lower `cil` and upper `ciu` values for my ± SEM (standard error of the mean).

`ciu` stands for "confidence interval lower;" `ciu` stands for "confidence interval upper."

Finally, here's the plot command

```
barplot2(hrtWt, beside = TRUE, main=c("Mice fed different amounts of copper in diet"
```

Now, draw a box around the graphic

```
box()
```

Whew!

What does your new graph look like? My graph is below (Fig. 11).

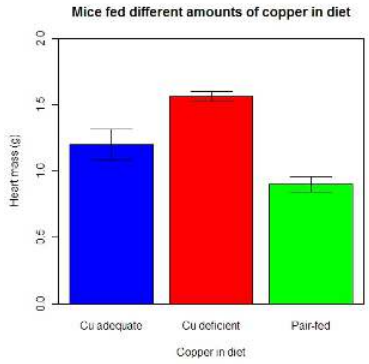


Figure 4.1.11: A bar chart made using `barplot2` .

This works, and the point is that once the script is written it easy to make small changes as you need in the future to make nice graphs.

If you are impatient like me, I like a GUI option, at least to start crafting the graph. The `Rcmdr` plugin `KMggplot2` provides a good set of tools to make bar charts with error bars. An even better option I think is to use a software package that is designed for graphics, at least simple graphics like a bar chart. I use SciDAVis and Veusz for simple graphs like pie charts and bar charts; much easier to control.

ggplot2 bar charts with error bars

Nevertheless, here's how to make a bar chart with error bars using `ggplot2` (Fig. 4.1.11). First, we need to create a statistics summary. The script printed here was modified from scripts at R Graph Cookbook website.

```
require(plyr)
summarySE <- function(data=NULL, measurevar, groupvars=NULL, na.rm=FALSE,
     conf.interval=.95, .drop=TRUE) {
length2 <- function (x, na.rm=FALSE) {
     if (na.rm) sum(!is.na(x))
     else length(x)
}

#returns a vector with N, mean, and sd
datac <- ddply(data,groupvars, .drop=.drop,
    .fun = function(xx,col){
      c(N=length2(xx[[col]],na.rm=na.rm),
      mean=mean(xx[[col]],na.rm=na.rm),
      sd=sd(xx[[col]],na.rm=na.rm)
     )
    },
    measurevar
   )

#Rename the "mean" column
datac <- rename(datac,c("mean"=measurevar))

#Calculate the standard error of the mean
datac$se <- datac$sd/sqrt(datac$N)

#Get confidence interval
ciMult <- qt(conf.interval/2 + .5, datac$N-1)
datac$ci <- datac$se*ciMult

return(datac)
}
```

Applying this function to the BMI dataset yields the following output.

```
sumBMI <- summarySE(BMI, measurevar="BMI", groupvars=c("Sex", "Smoke"));sumBMI

   Sex Smoke   N      BMI       sd        se       ci
1    F    No  23 26.83567 7.610271 1.5868511 3.290928
2    F   Yes  14 25.76133 4.658625 1.2450698 2.689810
3    M    No  10 26.35731 3.363575 1.0636557 2.406156
4    M   Yes  27 26.71879 4.675631 0.8998256 1.849618
```

Now we are ready to make the bar chart with error bars

```
ggplot(tgc,aes(x=Smoke,y=BMI,fill=Sex)) +
 geom_bar(position=position_dodge(),stat="identity",color="black") +
 geom_errorbar(aes(ymin=BMI,ymax=BMI+se),width=.2,position=position_dodge(.9))
```

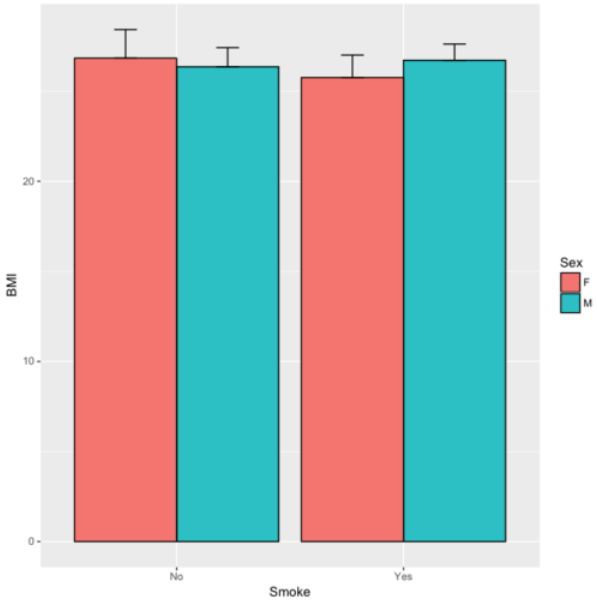


Figure 4.1.12: A barchart from `ggplot2` .

Questions

1. Why should you use box plots and not bar charts to display comparisons for a ratio scale variable between categories? Obtain a copy of the article by Streit and Gehlenbor 2014 — it's free! After reading, summarize the pro and cons for box plots over bar charts with error bars.

2. Enter the following data into R. The data are sulfate levels in water, parts per million.

```
type = c("Palolo Stream","Chaminade tap water", "Aquafina","Dasani")
sulfateppm =c(11, 14, 5, 12)
try = data.frame(type,sulfateppm)
byWater = tapply(try$sulfateppm,list(Group=try$type),mean)
```

Make a simple bar chart using the `boxplot2` function in `gplots` package.

3. Change the range of values on the vertical axis to 0, 20

4. Change the color of the bars from gray to blue

5. Add a label to the vertical axis, "Sulfates, ppm" (without the quotes)

6. Add a box around the graph.

Data set

Diet	Body	Heart	Liver
Adequate-Cu	320.1381	1.125037	10.259657
Adequate-Cu	329.6879	1.158982	9.843295
Adequate-Cu	327.9838	1.090374	9.855975
Adequate-Cu	334.6669	1.118183	9.942997
Adequate-Cu	338.3134	1.172636	9.860971
Adequate-Cu	345.4608	1.056183	8.88582
Adequate-Cu	343.089	1.081261	10.166647

Diet	Body	Heart	Liver
Adequate-Cu	328.3403	1.111278	10.124185
Adequate-Cu	324.9723	1.189194	10.158402
Adequate-Cu	325.2378	1.14715	9.939521
Deficient-Cu	195.5052	1.90973	7.907565
Deficient-Cu	182.7809	1.823672	8.430167
Deficient-Cu	184.3701	1.632249	7.619104
Deficient-Cu	193.7867	1.831765	8.742489
Deficient-Cu	180.0417	1.710367	7.975879
Deficient-Cu	208.5349	2.495623	8.652445
Deficient-Cu	182.3048	1.262053	7.257726
Deficient-Cu	203.0413	2.153639	8.081782
Deficient-Cu	193.3829	1.986028	7.807328
Deficient-Cu	195.0523	1.76975	8.297611
Pair-fed	211.0858	0.6911343	6.251177
Pair-fed	210.4041	0.6928067	7.696669
Pair-fed	208.5969	0.6911901	6.973803
Pair-fed	209.3333	0.7039211	6.629303
Pair-fed	208.8889	0.7077486	6.038704
Pair-fed	208.2994	0.7004535	6.606877
Pair-fed	209.4524	0.6915543	6.228888
Pair-fed	210.2699	0.6984497	6.638466
Pair-fed	208.8142	0.7214847	6.353705
Pair-fed	209.2977	0.6848656	6.536642

4.2: Histograms

Introduction

For displaying interval or continuously scaled data, a histogram (frequency or density distribution) is a useful graph to summarize patterns in data, and is commonly used to judge whether or not the sample distribution approximates a normal distribution. Three kind of histograms exist, depending on how the data are grouped and counted. Lump the data into a sequence of adjacent intervals or **bins** (aka **classes**), then count how many individuals have values that fall into one of the bins — the display is referred to as a **frequency histogram**. Sum up all of the frequencies or counts in the histogram and they add to the sample size. Convert from counts to percentages, then the heights of the bars are equal to the relative frequency (percentage) — the display is referred to as a **percentage histogram** (aka **relative frequency histogram**). Sum up all of the bin frequencies and they equal one (100%).

Figure 4.2.1 shows two frequency histograms of the distribution of ages for female (left panel) and male (right panel) runners at the 2013 Jamba Juice Banana 5K race in Honolulu, Hawaii (link to data set).

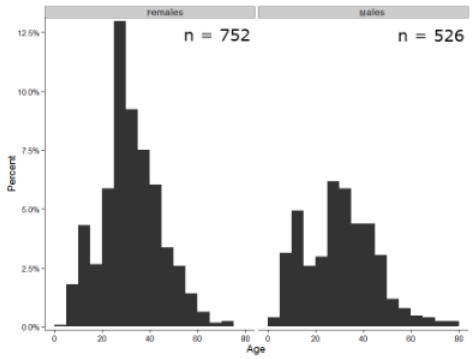


Figure 4.2.1: Histograms of age distribution of runners who completed the 2103 Jamba Juice 5K race.

The graphs in Fig. 4.2.1 were produced using R package `ggplot2` .

The third kind of histogram is referred to as a **probability density histogram**. The height of the bars are the **probability densities**, generally expressed as a decimal. The probability density is the bin probability divided by the bin width (size). The area of the bar gives the bin probability and the total area under the curve sums to one.

Which to choose? Both relative frequency histograms and density histograms convey similar messages because both “sum to one” (100%), i.e., bin width is the same across all intervals. Frequency histograms may have different bin widths; with more numerous observations, the bin width is larger than with cases with fewer observations.

Purpose of the histogram plot

The purpose of displaying the data is to give you or your readers a quick impression of the general distribution of the data. Thus, from our histogram one can see the range of the data and get a qualitative impression of the variability and the central tendency of the data.

Kernel density estimation

Kernel density estimation (KDE) is a **non-parametric** approach to estimate the probability distribution function. The “**kernel**” is a window function, where an interval of points is specified and another function is applied only to the points contained in the window. The function applied to the window is called the **bandwidth**. The **kernel smoothing function** then is applied to all of the data, resulting in something that looks like a histogram, but without the discreteness of the histogram.

The chief advantage of kernel smoothing over use of histograms is that histogram plots are sensitive to bin size, whereas KDE plot shapes are more consistent across different kernel algorithms and bandwidth choices.

Today, statisticians use kernel smoothing functions instead of histograms; these reduce the impact that binning has on histograms, although kernel smoothing still involves choices (Type of smoothing function? Default is Gaussian. Widths or bandwidths for smoothing? Varies, but the default is from the variance of the observations). Figure 4.2.2 shows a smoothed plot of the 752 age observations.

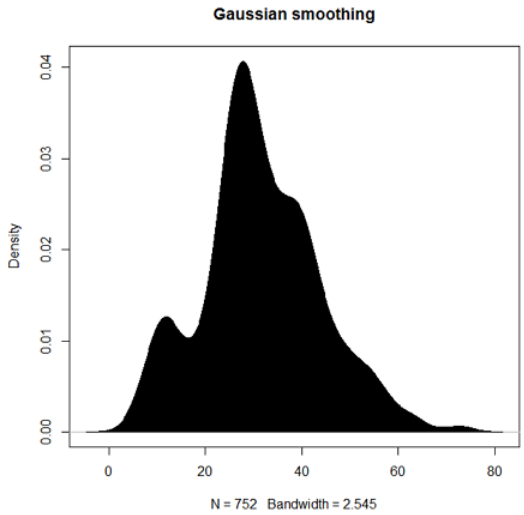


Figure 4.2.2: KDE plot of age distribution of female runners who completed the 2103 Jamba Juice 5K race in Honolulu.

Note:

Remember: the hashtag # preceding R code is used to provide comments and is not interpreted by R.

R commands typed at the R prompt were, in order:

```
d <- density(w) #w is a vector of the ages of the 752 females
plot(d, main="Gaussian smoothing")
polygon(d, col="black", border="black") #col and border are settings which allows you
```

Conclusion? A histogram is fine for most of our data. Based on comparing the histogram and the kernel smoothing graph I would reach the same conclusion about the data set. The data are right skewed, maybe kurtotic (peaked), and not normally distributed (see Ch 6.7).

Design criteria for a histogram

The X axis (horizontal axis) displays the units of the variable (e.g., age). The goal is to create a graph that displays the sample distribution. The short answer here is that there is no single choice you can make to always get a good histogram — in fact, statisticians now advise you to use a kernel function in place of histograms if the goal is to judge the distribution of samples.

For continuously distributed data the X-axis is divided into several intervals or bins:

1. The number of intervals depends (somewhat) on the sample size and (somewhat) on the range of values. Thus, the shape of the histogram is dependent on your choice of intervals: too many bins and the plot flattens and stretches to either end (over-smoothing); too few bins and the plot stacks up and the spread of points is restricted (under-smoothing). For both you lose the details of the histogram shape.
2. A general rule of thumb: try to have 10 to 15 different intervals. This number of intervals will generally give enough information.
3. For large sample size (N=1000 or more) you can use more intervals.

The intervals on the X-axis should be of equal size on the scale of measurement.

1. They will not necessarily have the same number of observations in each interval.
2. If you do this by hand you need to first determine the range of the data and then divide this number by the number of categories you want. This will give you the size of each category (e.g., range is 25; 25 / 10 = 2.5; each category would be 2.5 units).

For any given X category the Y-axis then is the number or frequency of individuals that are found within that particular X category.

Software

Rcmdr: Graphs → Histogram…

Accepting the defaults to a subset of the 5K data set, we get Fig. 4.2.3:

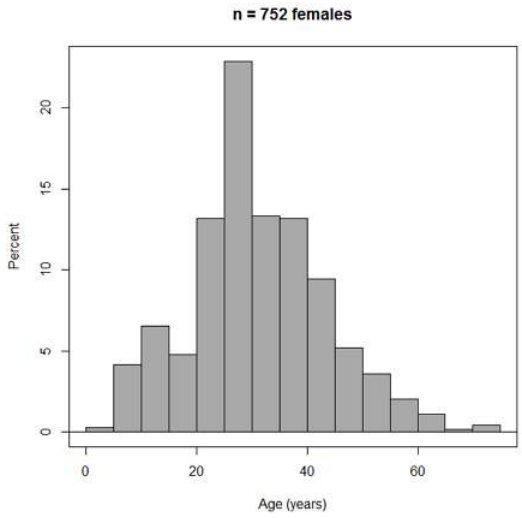


Figure 4.2.3: Histogram of 752 observations, **Sturge's rule** applied, default histogram.

The subset consisted of all females or n = 752 that entered and finished the 5K race with an official time.

R Commander plugin KMggplot2

Install the RcmdrPlugin.KMggplot2as you would any package in R. Start or restart Rcmdr and load the plugin by selecting **Rcmdr: Tools → Load Rcmdr plug-in(s)…** Once the plugin is installed select **Rcmdr: KMggplot2 → Histogram…** The popup menu provides a number of options to set to format the image. Settings for the next graph were No. of bins "Scott," font family "Bookman," Colour pattern "Set 1," Theme "theme_wsj2."

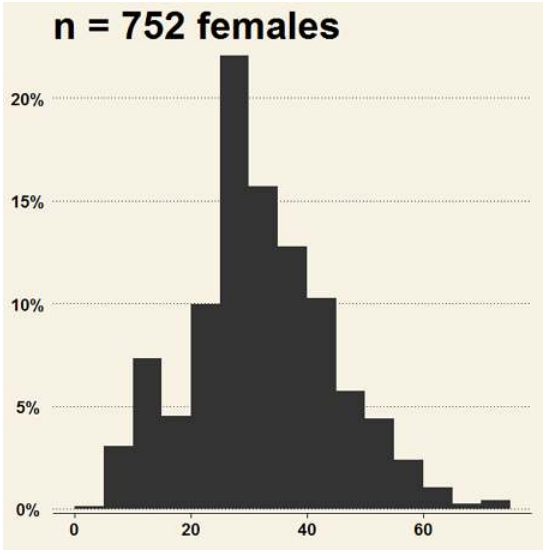


Figure 4.2.4: Histogram of 752 observations, Scott's rule applied, ggplot2 histogram.

Selecting the correct bin number

You may be saying to yourself, wow, am I confused. Why can't I just get a graph by clicking on some buttons? The simple explanation is that the software returns defaults, not finished products. It is your responsibility to know how to present the data. Now, the perfect graph is in the eye of the beholder, but as you gain experience, you will find that the default intervals in R bar graphs have too many categories (recall that histograms are constructed by lumping the data into a few categories, or bins, or

intervals, and counting the number of individuals per category => “density” or frequency). How many categories (intervals) should we use?

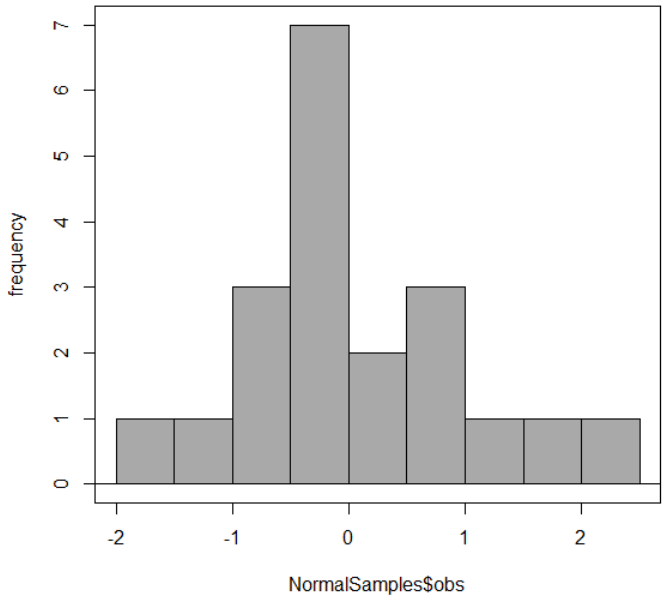


Figure 4.2.5: Default histogram with default bin size.

R’s default number for the intervals seems too much to me for this data set; too many categories with small frequencies. A better choice may be around 5 or 6. Change number of intervals to 5 (click Options, change from automatic to number of intervals = 5). Why 5? Experience is a guide; we can guess and look at the histograms produced.

Improving estimation of bin number

I gave you a rough rule of thumb. As you can imagine, there have been many attempts over the years to come up with a rational approach to selecting the intervals used to bin observations for histograms. The histogram function in Microsoft’s Excel (Data Analysis plug-in installed) uses the square root of the sample size as the default bin number. **Sturge’s rule** is commonly used, and the default choice in some statistical application software (e.g., Minitab, Prism, SPSS). **Scott’s** approach (Scott 2009), a modification to Sturge’s rule, is the default in the `ggplot()` function in the R graphics package (the Rcmdr plugin is `RcmdrPlugin.KMggplot2`). And still another choice, which uses interquartile range (IQR), was offered by **Freedman and Diacones** (1981). Scargle et al (1998) developed a method, **Bayesian blocks**, to obtain optimum binning for histograms of large data sets.

What is the correct number of intervals (bins) for histograms?

- Use the square root of the sample size, e.g., in this case the sample size $n = 20$ and $\sqrt{n} = 4.5$, round to 5.
- Follow Sturges’ rule (to get the suggested number of intervals for a histogram, let k = the number of intervals, and $k = 1 + 3.322\,(\log_{10} n)$, where n is the sample size.) I got $k = 5.32$, round to nearest whole number = 5.
- Another option was suggested by Freedman and Diacones (1981): find IQR for the set of observations and then the solution to the bin size is $k = 2 \cdot \mathrm{IQR} \cdot n^{-1/3}$, where n is the sample size.

Select Histogram Options, then enter intervals. Here’s the new graph using Sturge’s rule…

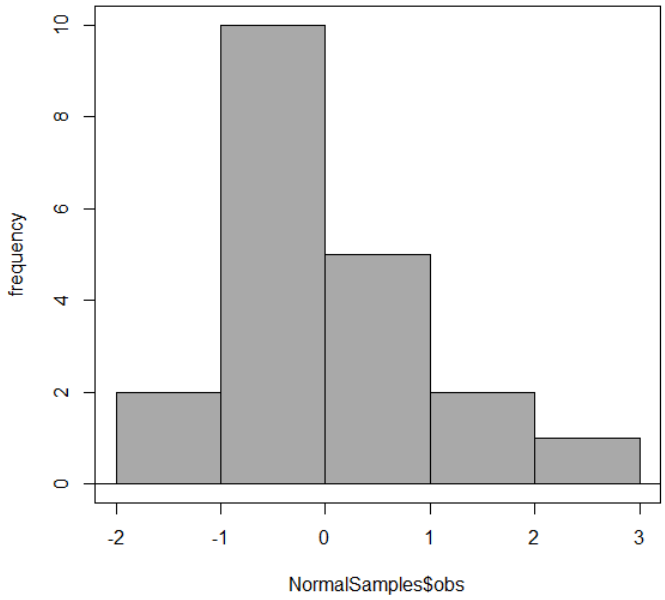


Figure 4.2.6: Default histogram, bin size set by Sturge's rule.

OK, it doesn't look much better. And of course, you'll just have to trust me on this — it is important to try to make the bin size appropriate given the range of values you have in order for the reader/viewer can judge the graphic correctly.

Questions

Example data set, comet tails and tea

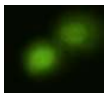


Figure 4.2.7: Examples of comet assay results.

The **Comet assay**, also called the single cell gel electrophoresis (SCGE) assay, is a sensitive technique to quantify DNA damage from single cells exposed to potentially mutagenic agents. Undamaged DNA will remain in the nucleus, while damaged DNA will migrate out of the nucleus (Figure 4.2.7). The basics of the method involve loading exposed cells immersed in low melting agarose as a thin layer onto a microscope slide, then imposing an electric field across the slide. By adding a DNA selective agent like Sybr Green, DNA can be visualized by fluorescent imaging techniques. A "tail" can be viewed: the greater the damage to DNA, the longer the tail. Several measures can be made, including the length of the tail, the percent of DNA in the tail, and a calculated measure referred to as the Olive Moment, which incorporates amount of DNA in the tail and tail length (Kumaravel et al 2009).

The data presented in Table 1 comes from an experiment in my lab; students grew rat lung cells (ATCC CCL-149), which were derived from type-2 like alveolar cells. The cells were then exposed to dilute copper solutions, extracts of hazel tea, or combinations of hazel tea and copper solution. Copper exposure leads to DNA damage; hazel tea is reported to have antioxidant properties (Thring et al 2011).

Data set, comet assay

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	10	9.7732	2.1501
Copper-Hazel	6	4.8381	0.9676
Copper-Hazel	6	3.981	0.836
Copper-Hazel	16	12.0911	2.9019

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	20	15.3543	3.9921
Copper-Hazel	33	33.5207	10.7266
Copper-Hazel	13	13.0936	2.8806
Copper-Hazel	17	26.8697	4.5679
Copper-Hazel	30	53.8844	10.238
Copper-Hazel	19	14.983	3.7458
Copper	11	10.5293	2.1059
Copper	13	12.5298	2.506
Copper	27	38.7357	6.9724
Copper	10	10.0238	1.9045
Copper	12	12.8428	2.5686
Copper	22	32.9746	5.2759
Copper	14	13.7666	2.6157
Copper	15	18.2663	3.8359
Copper	7	10.2393	1.9455
Copper	29	22.6612	7.9314
Hazel	8	5.6897	1.3086
Hazel	15	23.3931	2.8072
Hazel	5	2.7021	0.5674
Hazel	16	22.519	3.1527
Hazel	3	1.9354	0.271
Hazel	10	5.6947	1.3098
Hazel	2	1.4199	0.2272
Hazel	20	29.9353	4.4903
Hazel	6	3.357	0.6714
Hazel	3	1.2528	0.2506

Rat lung cells treated with Hazel tea extract and exposed to copper metal. Tail refers to length of the comet tail, TailPercent is percent DNA damage in tail, and Olive moment refer's to Olive (1990), defined as the fraction of DNA in the tail times the tail length.

Copy the table into a data frame.

1. Create histograms for tail, tail percent, and olive moment
   - Change bin size
2. Repeat, but with a kernel function.
3. Looking at the results from question 1 and 2, how “normal” (i.e., equally distributed around the middle) do the distributions look to you?

4. Plot means to compare Tail, Tail percent, and olive moment. Do you see any evidence to conclude that one of the teas protects against DNA damage induced by copper exposure?

4.3: Box plots

Introduction

Box plots, also called **whisker plots**, should be your routine choice for exploring ratio scale data. Like bar charts, box plots are used to compare ratio scale data collected for two or more groups. Box plots serve the same purpose as bar charts with error bars, but box plots provide more information.

Purpose and design criteria

Box plots are useful tool for getting a sense of central tendency and spread of data. These types of plots are useful diagnostic plots. Use them during initial stages of data analyses. All summary features of box plots are based on ranks (not sums). So, they are less sensitive to extreme values (outliers). Box plots reveal asymmetry. Standard deviations are symmetric.

The median splits each batch of numbers in half (center line). The "hinge" (median value) splits the remaining halves in half again (the quartiles). The first, second (median), and third quartiles describes the interquartile range, or IQR, 75% of the data (Fig. 4.3.1). Outlier points can be identified, for example, with an asterisk or by **id number** (Fig. 4.3.1).

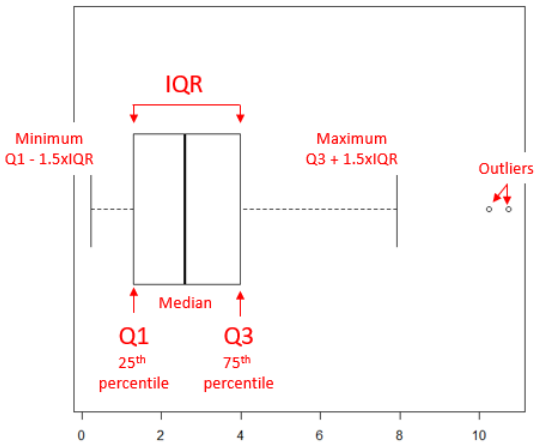


Figure 4.3.1: A box plot. Elements of box plot labeled.

We'll use the data set described in the previous section, so if you have not already done so, get the data from Table 1, Chapter 4.2 into your R software.

Note:

See Chapter 4.10 — Graph software for additional box plot examples, but made with different R packages or software apps.

R Code

Command line

We'll provide code for the base graph shown in Figure 4.3.2*A*. At the R prompt, type

```
boxplot(OliveMoment~Treatment)
```

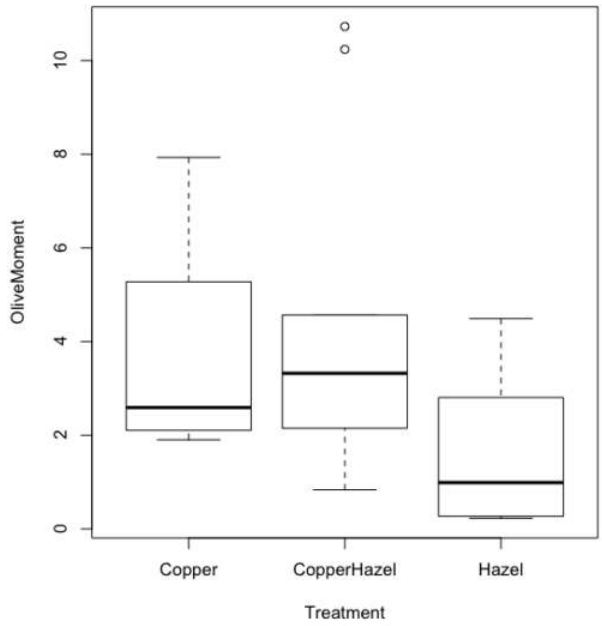


Figure 4.3.2A: Box plot, default graph in base package.

Boxplot is a common function offered in several packages. In the base installation of R, the function is `boxplot()` . The `car` package, which is installed as part of R Commander installation, includes `Boxplot()` , which is a "wrapper function" for `boxplot()` . Note the difference: `base` package is all lower case, `car` package the "B" is uppercase. One difference, base `boxplot()` permits horizontal orientation of the plot (Fig. 4.3.2B).

Note:

Wrapper functions are code that links to another function, perhaps simplifying working with that function.

```
boxplot(OliveMoment ~ Treatment,  horizontal=TRUE, col="steelblue")
```

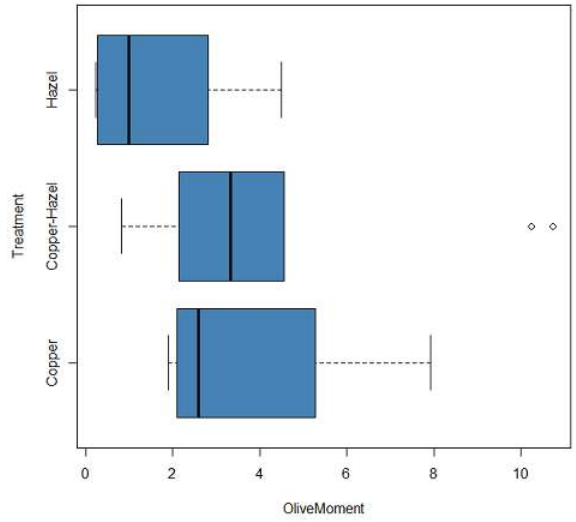


Figure 4.3.2B: Same graph, but with color and made horizontal; `boxplot()` , default graph in base package.

Base package `boxplot()` has additional features and options compared to `Boxplot()` in the car package. i.e., not all `barcode()` options are wrapped. For example, I had more success adding original points to `boxplot()` graph (Fig.

4.3.2*C*) following the function call with `stripchart()` .

```
stripchart(OliveMoment ~ Treatment, method = "overplot", pch = 19, add = TRUE)
```

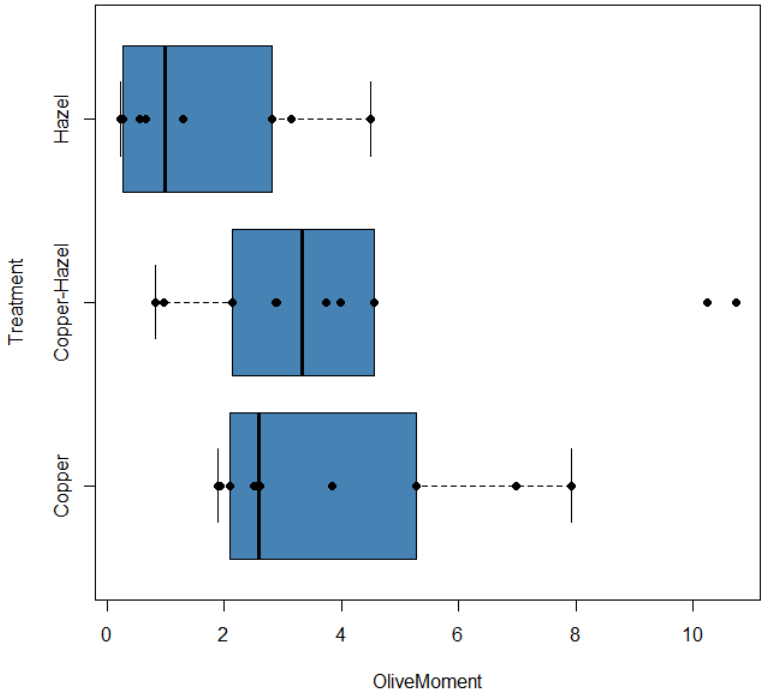


Figure 4.3.2*C*: Same graph, added original points; `boxplot()` , default graph in base package.

Note:

`boxplot` and `stripchart` functions are part of `ggplot2` package, part of `tidyverse` , and easily used to generate graphs like Fig. 4.3.2*B* and Fig. 4.3.2*C*. The `overplot` option was used to **jitter** points to avoid **overplotting**. See below: **Apply tidyverse-view to enhance look of boxplot graphic** and Fig. 4.3.9.

Jittering adds random noise to points, which helps view the data better if many points are clustered together. Note however that jitter would add noise to the plot — if the objective is to show an association between two variables, jitter will reduce the apparent association, perhaps even compromising the intent of the graph. **Beeswarm** also can be used to better visualize clustered points, but uses a nonrandom algorithm to plot points.

Rcmdr: Graph → Boxplot…

Select the response variable, then click on the “Plot by:” button

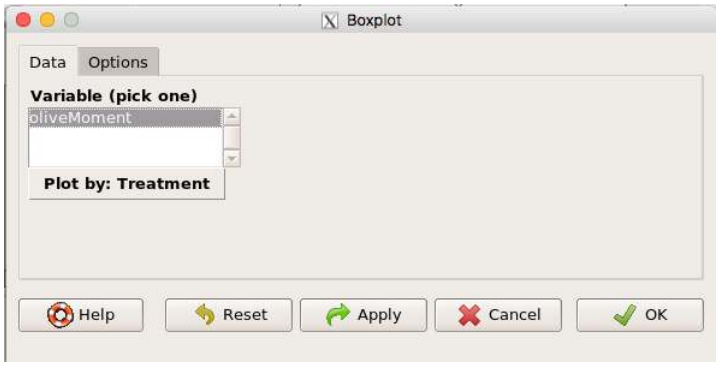


Figure 4.3.3: Boxplot popup menu in R Commander. Select the response variable and set the “Plot by:” option.

Next, select the Groups (Factor) variables (Fig. 4.3.4). Click OK to proceed.

Figure 4.3.4: Select the group variable.

Back to the Box Plot menu, click “Options” tab to add details to the plot, including a graph title and how outliers are noted (Fig. 4.3.5),

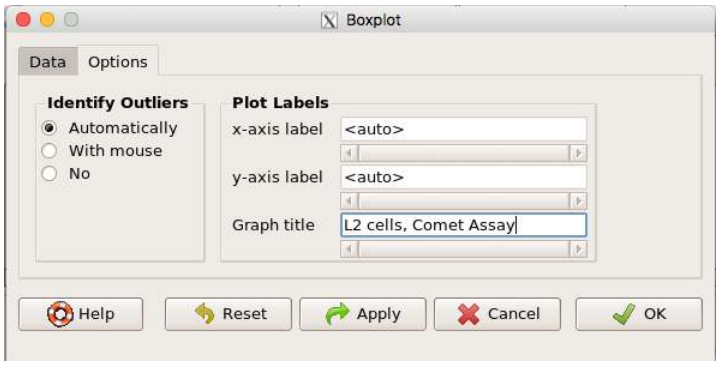


Figure 4.3.5: Options tab of boxplot popup. Enter axes labels and a title.

And here is the resulting box plot (Fig. 4.3.6)

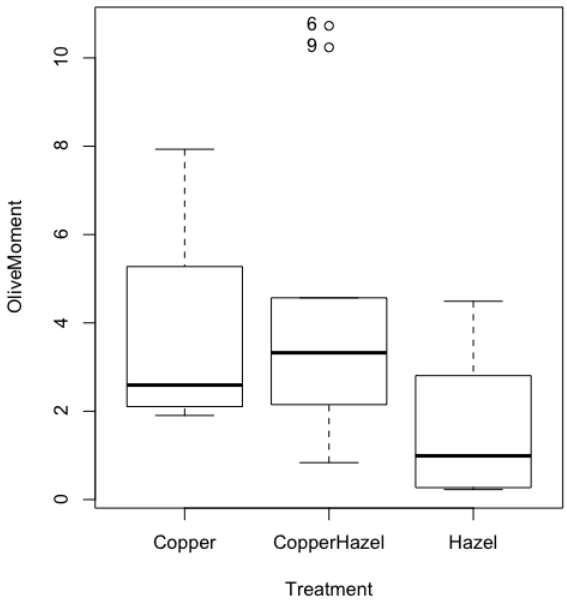


Figure 4.3.6: Resulting box plot from car package implemented in R Commander. Outliers are identified by row id number.

The graph is functional, if not particularly compelling. The data set was “olive moments” from Comet Assays of an immortalized rat lung cell line exposed to dilute copper solution (Cu), Hazel tea (Hazel), or Hazel & Copper solution.

Apply Tidyverse-view to enhance look of boxplot graphic

Load the `ggplot2` package via the Rcmdr plugin to add options to your graph. As a reminder, to install Rcmdr plugins you must first download and install them from an R mirror like any other package, then load the plugin via **Rcmdr Tools → Load Rcmdr plug-in(s)…** (Fig. 4.3.7, Fig. 4.3.8).

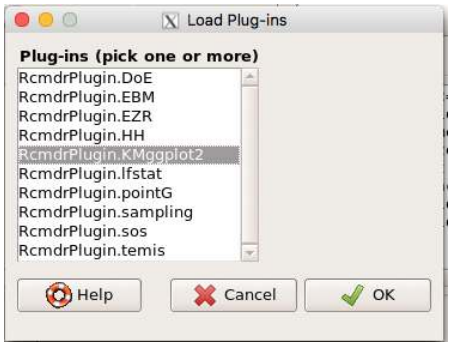


Figure 4.3.7: Screenshot of Load Rcmdr plug-ins menu, `ggplot2` selected. Click OK to proceed (see Fig. PageIndex{8}\).

Figure 4.3.8: To complete installation of the plug-in, restart R Commander.

Significant improvement, albeit with an “eye of the beholder” caveat, can be made over the base package. For example, `ggplot2` provides additional themes to improve on the basic box plot. Figure 4.3.9 shows the options available in the Rcmdr plugin `KMggplot2`, and the default box plot is shown in Figure 4.3.10.

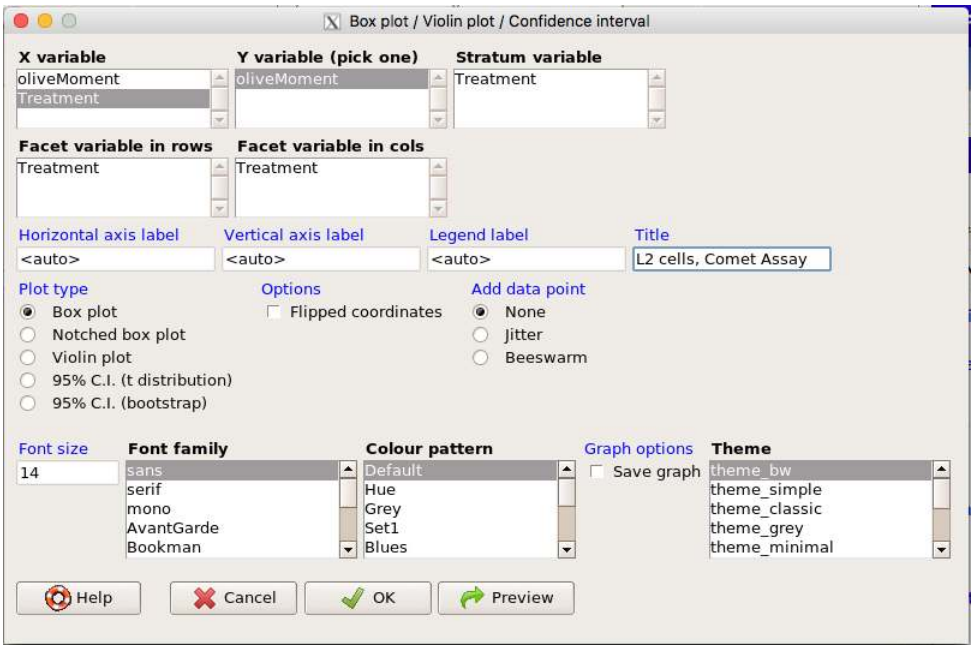


Figure 4.3.9: Menu of KMggplot2. A title was added, all else remained set to defaults.

The next series of plots explore available formats for the charts.

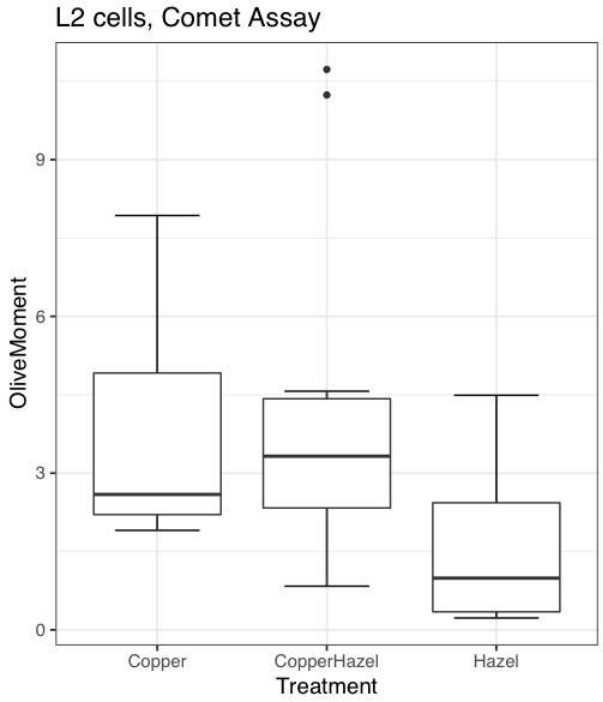


Figure 4.3.10: Default box plot from KMggplot2.

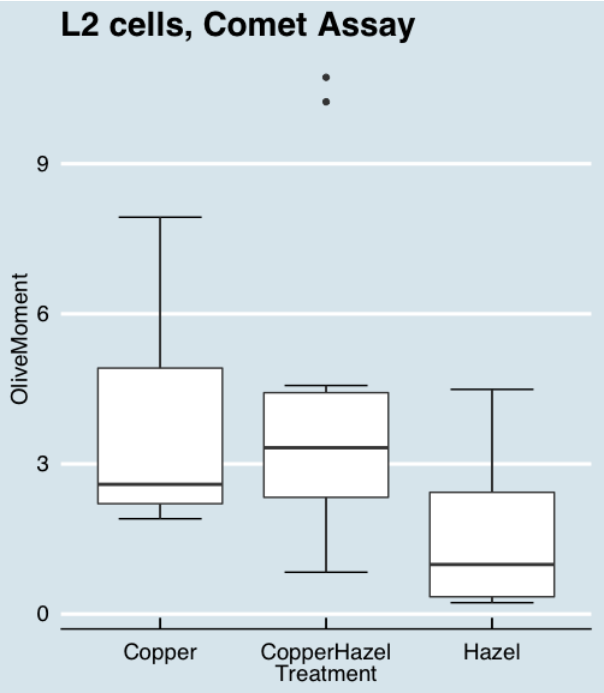


Figure 4.3.11: “Economist” theme box plot from KMggplot2.

And finally, since the box plot is often used to explore data sets, some recommend including the actual data points on a box plot to facilitate pattern recognition. This can be accomplished in the `KMggplot2` plugin by checking “**Jitter**” under the **Add data points** option (see Fig. 4.3.9). Jitter helps to visualize overlapping points at the expense of accurate representation. I also selected the Tufte theme, which results in the image displayed in Figure 4.3.12

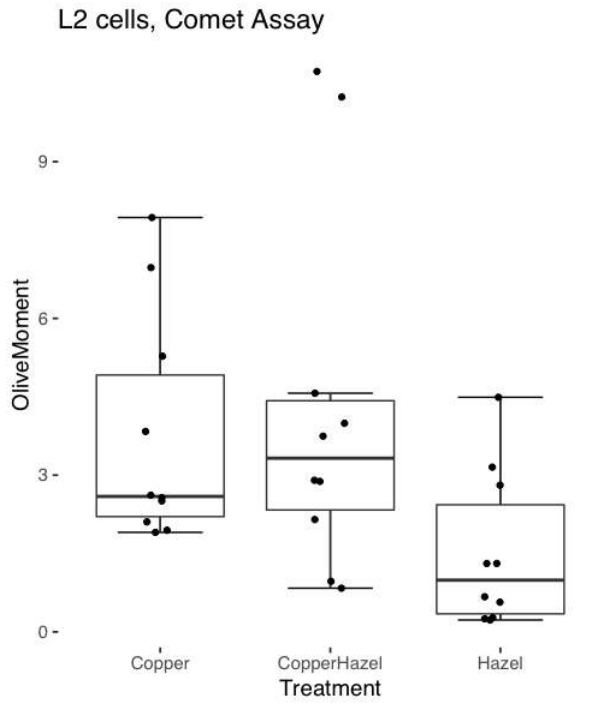


Figure 4.3.12: Tufte theme and data points added to the box plot.

Conclusions

As part of your move from the world of Microsoft Excel graphics to graphs recommended by statisticians, the box plot is used to replace the bar charts plus error bars that you may have learned in previous classes. The second conclusion? I presented a number of versions of the same graph, differing only by style. Pick a style of graphics and be consistent.

Questions

1. Why is a box plot preferred over a bar chart for ratio scale data, even if an appropriate error bar is included?
2. With your comet data (Table 1, Chapter 4.2), explore the different themes available in the box plot commands available to you in Rcmdr. Which theme do you prefer and why?

4.4: Mosaic plots

Introduction

Mosaic plots are used to display associations among categorical variables. e.g., from a contingency table analysis. Like pie charts, mosaic plots and tree plots (next chapter) are used to show part-to-whole associations. Mosaic plots are simple versions of heat maps (next chapter). Used appropriately, mosaic plots may be useful to show relationships. However, as with pie charts and bar charts, care needs to be taken to avoid their overuse; a mosaic plot works for a few categories, but quickly loses clarity as numbers of categories increase.

In addition to the function `mosaicplot()` in the base R package, there are a number of packages in R that will allow you to make these kinds of plots; depending on the analyses we are doing we may use any one of three Rcmdr plugins: `RcmdrPlugin.mosaic` (depreciated), `RcmdrPlugin.KMggplot2` , or `RcmdrPlugin.EBM` .

Example data

Table 4.4.1. Home wins record of American and National Leagues baseball teams at home and away midway through 2016 season

	No	Yes
AL	10	5
NL	7	8

The configuration of major league baseball (MLB) parks differ from city to city. For example, Boston's American League (AL) Fenway Park has the 30-feet tall "Green Monster" fence in left field and a short distance of only 302 feet along the foul line to right field fence. For comparison, in Globe Life Park in Arlington, TX the distance along the foul lines is 332 feet for left field and 325 feet for right field. So, it suggests that teams may benefit from playing 81 games at their home stadium. To test this hypothesis I selected Win-Loss records of the 30 teams at the midway point of the 2016 season. Data are shown in Table 4.4.1.

mosaicplot() in R base

The function `mosaicplot()` is included in the base install of R. The following code is one way to directly enter contingency table data like that from Table 1.

```
myMatrix <- matrix(c(10,  5,  7,  8),  nrow = 2, ncol = 2, byrow = TRUE)
dimnames(myMatrix) <- list(c("AL", "NL"), c("No","Yes"))
myTable <- as.table(myMatrix); myTable
mosaicplot(myTable, color=2:3)
```

The simple plot is shown in Figure 4.4.1. color = "2" is red, color = "3" is green.

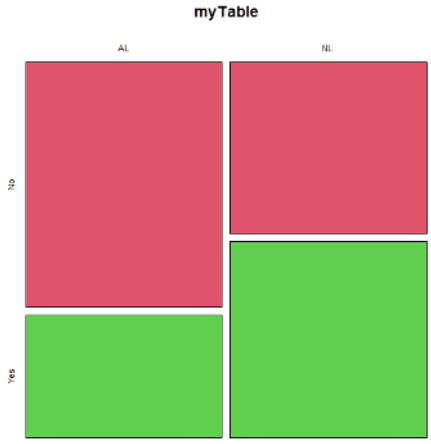


Figure 4.4.1: Mosaic plot made with basic function `mosaicplot()` .

mosaic plot from EBM plugin

A good option in `Rcmdr` is to use the "evidence-based-medicine" or "EBM" plug-in for Rcmdr (`RcmdrPlugin.EBM`). This plugin generates a really nice mosaic plot for 2 × 2 tables.

After loading the EBM plugin, restart `Rcmdr` , then select EBM from the menu bar and choose to "Enter two-way table…"

Figure 4.4.2: First steps to make mosaic plot in R Commander EBM plug-in.

Complete the data entry for the table as shown in the image below. After entering the values, click the OK button.

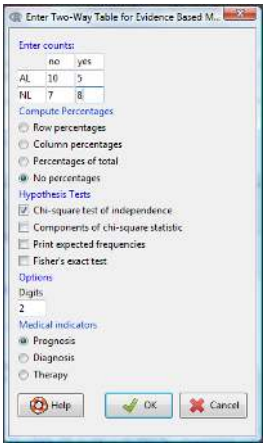


Figure 4.4.3: Next steps to make mosaic plot in R Commander EBM plug-in.

Along with the requested statistics, a mosaic plot will appear in a pop-up window.

Figure 4.4.4: Mosaic plot made from R Commander EBM plug-in.

mosaic-like plot KMggplot2 plugin

The `KMggplot2` plugin for Rcmdr will also generate a mosaic-like plot. After loading the `KMggplot2` plugin, restart Rcmdr, then load a data set with the table (e.g., MLB data in Table 4.4.1). Next, from within the `KMggplot2` menu select, "Bar chart for discrete variables…"

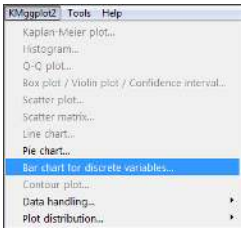


Figure 4.4.5: First steps to make mosaic plot in R Commander KMggplot2 plug-in.

From the bar chart context menu make your selections. Note that this function has many options for formatting, so play around with these to make the graph the way you prefer.

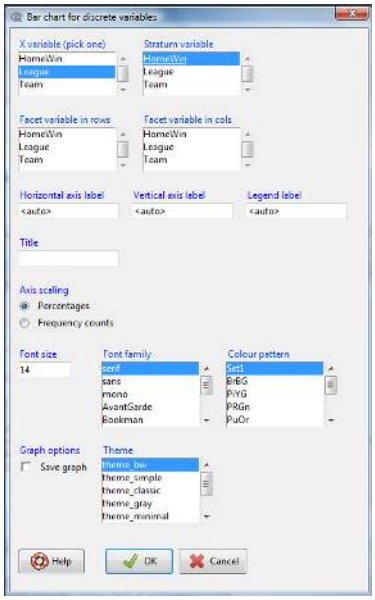


Figure 4.4.6: Next steps to make mosaic plot in R Commander KMggplot2 plug-in.

And here is the resulting mosaic-like plot from `KMggplot2` .

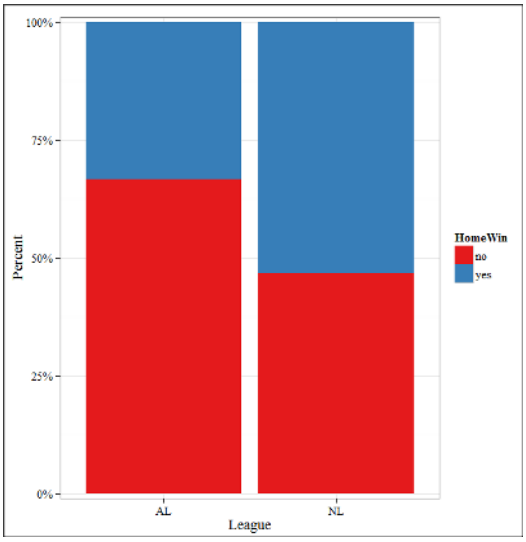


Figure 4.4.7: Mosaic-like plot made from R Commander KMggplot2 plug-in.

Depreciated material

As of summer 2020, `Rcmdrplugin.mosaic` is depreciated. While you can install the archived version, it is not recommended. Therefore, this material is left as is but for information purposes only. For a simple mosaic plot in `Rcmdr` I recommend working with the `RcmdrPlugin.EBM` .

Download the `RcmdrPlugin.mosaic` package, start `Rcmdr` , then navigate to `Tools` and choose **Load Rcmdr plug-in(s)…** Select **Rcmdrplugin.mosaic** (Fig. 4.4.8), then restart `Rcmdr` (Fig. 4.4.9). The plugin adds mosaic plot to the regular Graphics menu of Rcmdr.

Figure 4.4.8: Screenshot of popup menu from Rcmdr with mosaic plugin selected.

Figure 4.4.9: After clicking OK (Fig. 4.4.8), click Yes to restart Rcmdr. The plugin will then be available.

Load a data set with 2X2 arranged data, or create the variables yourself (Yikes, 30 rows!). The mosaic plugin requires that you submit data in a table format. We can check whether our data are currently in that format. At the R prompt type

```
is.table(MLB)
```

And R will return

```
[1] FALSE
```

(To be complete, confirm that the data set is a data.frame: `is.data.frame(MLB)` .)

You will need a table before proceeding with the mosaic plug-in. Then create a table using a command like the one shown below.

```
MLBTable <- xtabs(~League+HomeWin, data=MLB)
```

Once the table is ready, select “mosaic or assoc plot” from the Rcmdr Graphics menu (Fig. 4.4.10)

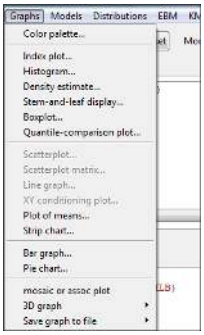


Figure 4.4.10: How to access the mosaic plot in R Commander.

A small window will pop up that will allow you to select the table of data you just created (Fig. 4.4.11). Note that you may need to hunt around your desktop to find this menu! Select the table (in this example, “MLBTable”), then click on “Create plot” button.

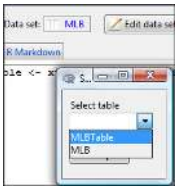


Figure 4.4.11: Screenshot of popup menu in mosaic plugin in R Commander.

R Note: The popup from the mosaic menu shown in Fig. 4.4.11 will also display the `data.frame MLB` . If you mistakenly select the `dataframe MLB` , you’ll get an error message in Rcmdr (Fig. 4.4.12). The plugin behaves erratically if you select MLB: On my computer, the function hangs and requires restarting R.

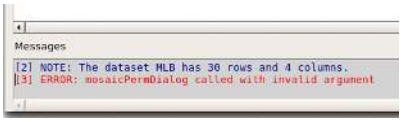


Figure 4.4.12: Error message as result of selecting a dataframe for use in mosaic plugin.

After you select the table, two additional windows will pop up: on the left (Fig. 4.4.13) is the context menu to change characteristics of the mosaic plot; on the right (not shown) will be a mosaic plot itself in default greyscale colors.

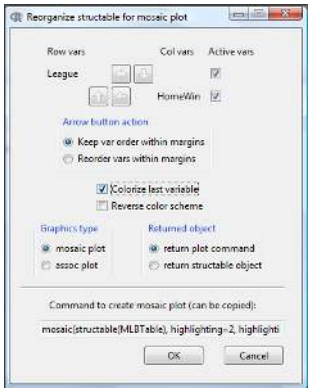


Figure 4.4.13: Options for the mosaic plot.

At a minimum, change the plot from greyscale to a colorized version by checking the box next to the “Colorize last variable” option. The new plot is shown in Figure 4.4.14

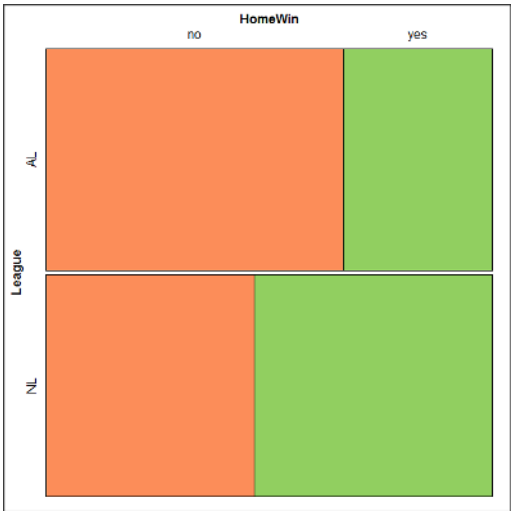


Figure 4.4.14: Our new mosaic plot.

OK. Take a moment and look at the plot. What conclusions can be made about our hypothesis — are there any differences between the leagues for home versus road Wins-Loss records?

By default the mosaic command copies the command to the R window. You can change the graph by taking advantage of the options in the brewer palette. Here's the command for the mosaic image above.

```
mosaic(structable(MLBTable), highlighting=2, highlighting_fill=brewer.pal.ext(2,"RdYlGn"))
```

Change the options in the brackets following " `brewer.pal.ext` ." For example, replace `RdYlGn` with `Blues` to make a plot that looks like the following:

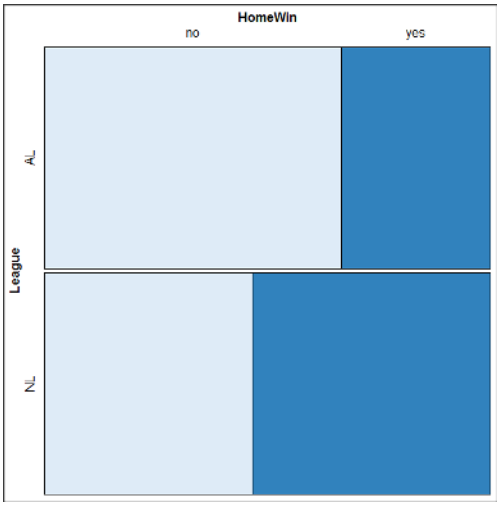


Figure 4.4.15: Mosaic plot with changed color scheme.

The colors are selected from the Rcolorbrewer package. For more, see this blog for starters.

Questions

1. Most US states have laws that dictate pre-employment drug testing for job candidates; Interestingly, states are increasingly legalizing marijuana use. Data for states plus District of Columbia are presented in the table. Make a mosaic plot of the table.

Table 4.4.2. Status of pre-employment drug testing by state.

	Marijuana use legal	Marijuana use not leg
Yes	19	12
No	14	6

Data adopted from https://www.paycor.com/resource-center/pre-employment-drug-testing-laws-by-state

4.5: Scatter plots

Introduction

Scatter plots, also called **scatter diagrams**, **scatterplots**, or **XY plots**, display associations between two quantitative, ratio-scaled variables. Each point in the graph is identified by two values: its X value and its Y value. The horizontal axis is used to display the dispersion of the X variable, while the vertical axis displays the dispersion of the Y variable.

The graphs we just looked at with Tufte's examples of **Anscombe's quartet** data were scatter plots (Chapter 4 – How to report statistics).

Here's another example of a scatter plot using data from Francis Galton, as contained in the R package `HistData` .

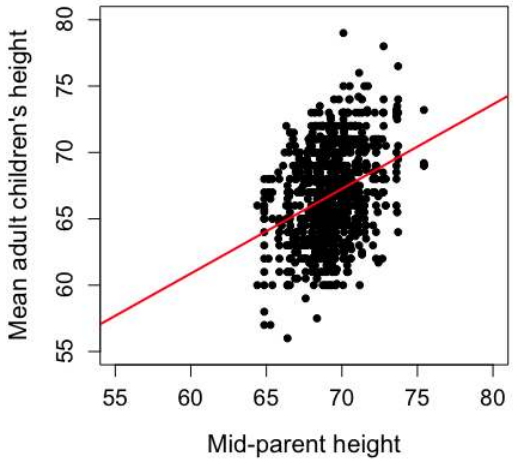


Figure 4.5.1: Scatterplot of mid-parent (horiztonal axis) and their adult children's (vertical axis) height, in inches. Data from Galton's 1885 paper, "Regression towards mediocrity in hereditary stature." The red line is the linear regression fitted line, or "trend" line, which is interpreted in this case as the heritability of height.

The commands I used to make this plot were

```
library(HistData)
data(GaltonFamilies, package="HistData")
attach(GaltonFamilies)

plot(childHeight~midparentHeight, xlab="Mid-parent height", ylab="Mean adult childre
abline(lm(childHeight~midparentHeight), col="red", lwd=2)
```

I forced the plot function to use the same range of values, set by providing values for `xlim` and `ylim` ; the default values of the plot command picks a range of data that fits each variable independently. Thus, the default X axis values ranged from 64 to 76 and the Y variable values ranged from 55 to 80. This has the effect of shifting the data, reducing the amount of white space, which a naïve reading of Tufte would suggest is a good idea, but at the expense of allowing the reader to see what would be the main point of the graph: that the children are, on average, shorter than the parents, mean height = 67 vs. 69 inches, respectively. Therefore, Galton's title begins with the word "regression," as in the definition of regression as a "return to a former … state" (Oxford Dictionary).

For completeness, `cex` sets the size of the points (default = 1), and therefore `cex.axis` and `cex.lab` apply size changes to the axes and labels, respectively; `pch` refers to the graph elements or plotting characters, further discussed below; `lm()` is a call to the linear model function; `col` refers to color.

Figure 4.5.2 shows the same plot, but without attention to the axis scales.

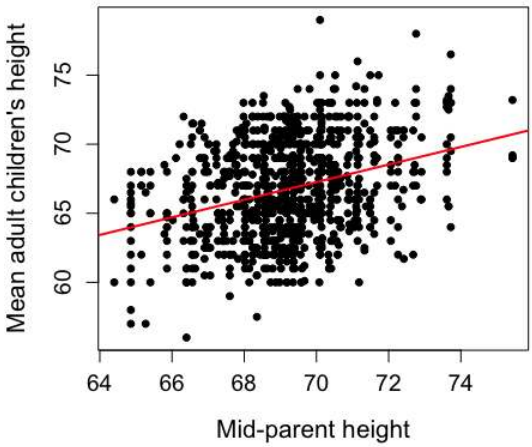


Figure 4.5.2: Same plot as Figure 4.5.1, but with default settings for axis scales.

Take a moment to compare the graphs in Figures 4.5.1 and 4.5.2. Setting the scales equal allows you to see that the mid-parent heights were less variable, between 65 and 75 inches, than the mean children height, which ranged from 55 to 80 inches.

And another example, Figure 4.5.3. This plot is from the `ggplot2()` function and was generated from within R Commander's `KMggplot2` plug-in.

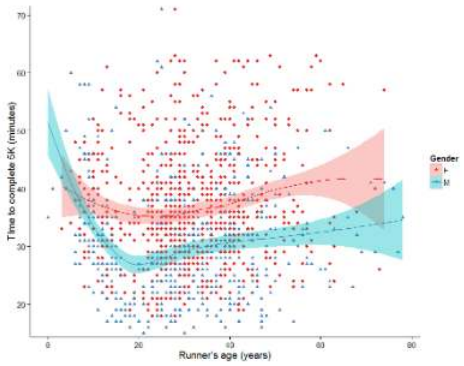


Figure 4.5.3: Finishing times in minutes of 1278 runners by age and gender at the 2013 Jamba Juice Banana 5K in Honolulu, Hawaii. **Loess smoothing functions** by groups of female (red) and male (blue) runners are plotted along with 95% confidence intervals.

Figure 4.5.3 is a busy plot. Because there were so many data points, it is challenging to view any discernible pattern, unlike the Figure 4.5.1 and 4.5.2 plots, which featured less data. Use of the Loess smoothing function, a transformation of the data to reduce data "noise" to reveal a continuous function, helps reveal patterns in the data:

1. across most ages, men completed the 5K faster than did females and
2. there was an inverse, nonlinear association between runner's age and time to complete the 5K race.

Take a look at the X-axis. Some runners' ages were reported as less than 5 years old (trace the points down to the axis to confirm), and yet many of these youngsters were completing the 5K race in less than 30 minutes. That's under a 10-minute mile pace. What might be some explanations for how pre-schoolers could be running so fast?

Design criteria

As in all plotting, maximize information to background. Keep white space minimal and avoid distorting relationships. Some things to consider:

1. keep axes same length
2. do not connect the dots UNLESS you have a continuous function
3. do not draw a trend line UNLESS you are implying causation

Scatter plots in R

We have many options in R to generate scatter plots. We have already demonstrated use of `plot()` to make scatter plots. Here we introduce how to generate the plot in R Commander.

Rcmdr: Graphs → Scatterplot…

Rcmdr uses the `scatterplot` function from the `car` package. In recent versions of R Commander the available options for the scatterplot command are divided into two menu tabs, **Data** and **Options**, shown in Figure 4.5.4 and Figure 4.5.5.

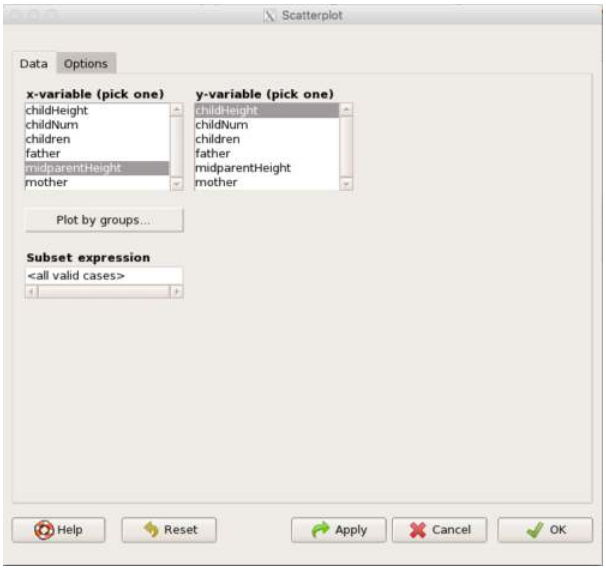


Figure 4.5.4: First menu popup in R Commander Scatterplot command, Rcmdr ver. 2.2-3.

Select X and Y variables, choose **Plot by groups** if multiple grounds are included, e.g., male, female, then click **Options** tab to complete.

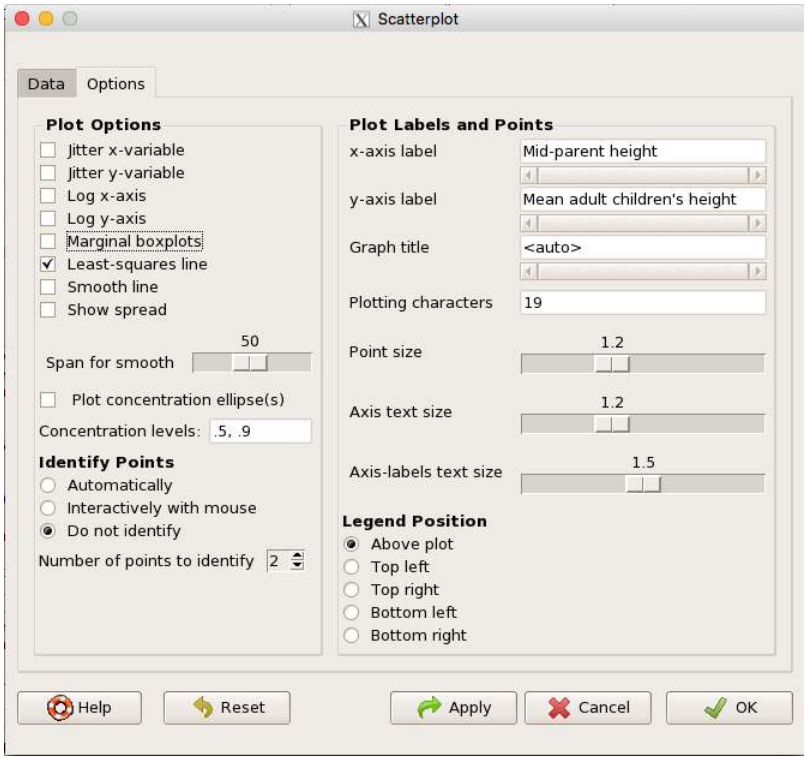


Figure 4.5.5: Second menu popup in R Commander scatterplot command., Rcmdr ver. 2.2-3.

Set graph options, including axis labels and size of the points.

Note 1:

There are lots of boxes to check and uncheck. Start by unchecking all of the **Options** and do update the axis labels. You can also manipulate the plot "points," which R refers to as plotting characters (abbreviated `pch` in plotting commands). The "Plotting characters" box is shown as <auto>, which is an open circle. You can change this to one of 26 different characters by typing in a number between 0 and 25. The default used in Rcmdr scatterplot is "1" for open circle. I typically use "19" for a solid circle.

Here is another example using the default settings in `scatterplot()` function in the car package, now the default scatter plot command via R Commander (Fig. 4.5.6), along with the same graph, but modified to improve the look and usefulness of the graph (Fig. 4.5.7). The data set was `Puromycin` in the package `datasets` .

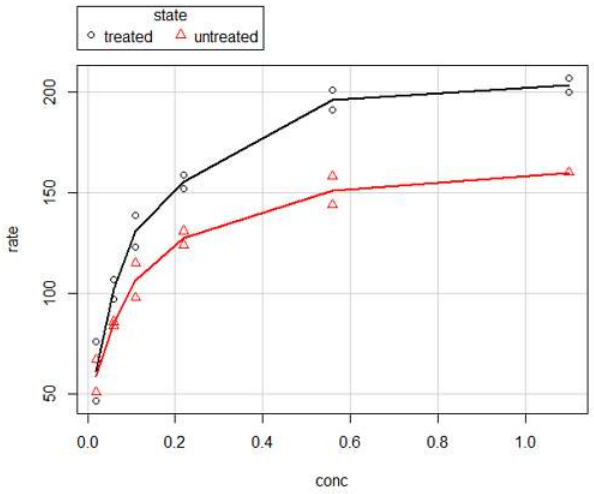


Figure 4.5.6: Default scatterplot, package car, from R Commander, version 2.2-4.

Grid lines in graphs should be avoided unless you intend to draw attention to values of particular data points. I prefer to position the figure legend within the frame of the graph, e.g., the open are at the bottom right of the graph. Modified graph shown in Figure 4.5.7.

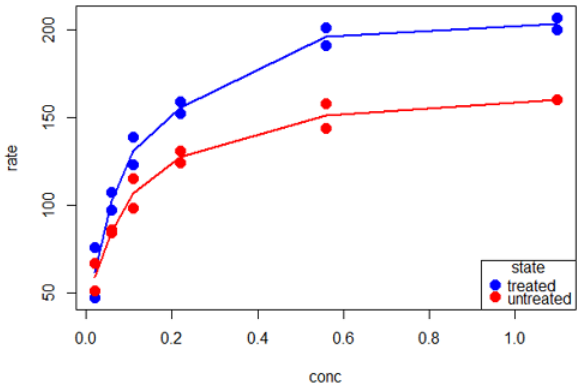


Figure 4.5.7: Modified scatterplot, same data from Figure 4.5.6.

R commands used to make the scatter plot in Figure 4.5.7 were

```
scatterplot(rate~conc|state, col=c("blue", "red"), cex=1.5, pch=c(19,19),
bty="n", reg=FALSE, grid=FALSE, legend.coords="bottomright")
```

A comment about graph elements in R

In some ways R is too rich in options for making graphs. There are the plot functions in the base package, there's lattice and `ggplot2` which provide many options for graphics, and more. The advice is to start slowly and explore. For example, you

might want to create something like Figure 4.5.8, which displays R's plotting characters and the number you would invoke to retrieve that plotting character.

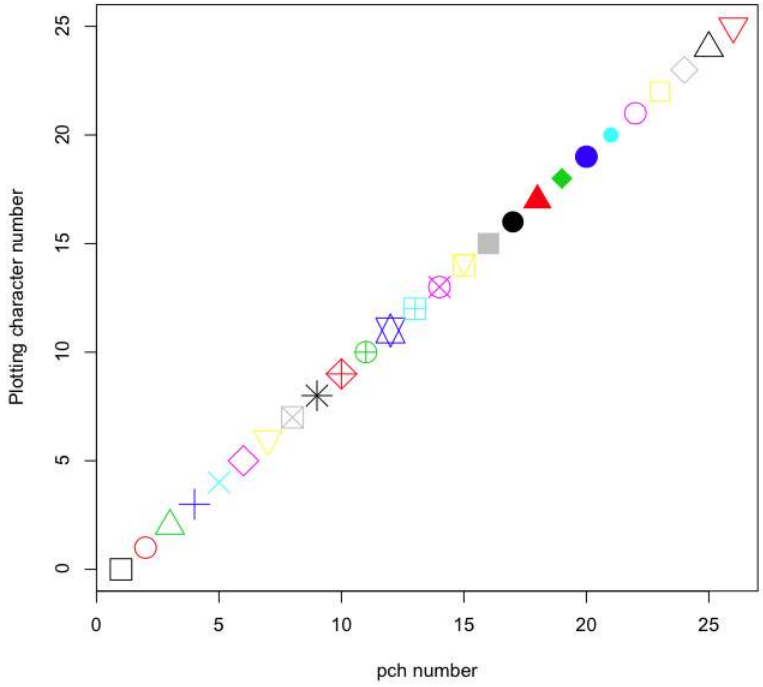


Figure 4.5.8: R plotting characters pch = 1 – 25, along with examples of color.

Note 2:

To see available colors, at the R prompt type

```
colors()
```

which returns 667 different colors by name, from

```
  [1] "white"                "aliceblue"            "antiquewhite"
```

to

```
[655] "yellow3"              "yellow4"              "yellowgreen"
```

Note 3:

There's a lot more to R plotting. For example, you are not limited to just 25 possible characters. R can print any of the ASCII characters 32:127 or from the extended ASCII code 128:255. See Wikipedia to see the listing of ASCII characters.

Note 4:

You can change the size of the plotting character with "cex."

Here's the R code used to generate the graph in Figure 4.5.8. Remember, any line beginning with # is a comment line, not an R command.

```
#create a vector with 26 numbers, from 0 to 25
stuff <-c(0:25)
plot(stuff, pch=c(32:58), cex = 2.5, col = c(1:26), 'xlab' = "pch number", 'ylab' = '
```

Is it “scatter plot” or “scatterplot”?

Spelling matters, of course, and yet there are many words for which the correct spelling seems to be like “beauty,” it is in the eye of the beholder. Scatter plot is one of these — is it one word or two?

And I’m not just talking about the differences between British and American English for many words, as listed at web sites like http://www.tysto.com/uk-us-spelling-list.html. Scatter plot is one of these terms: you’ll find it spelled as “scatterplot” or as “scatter plot,” in the dictionary (e.g., Oxford English dictionary), with no guidance to choose between them.

The spell checkers in Microsoft Office and Google Docs do not flag “scatterplot” as incorrect, but the spell checker in LibreOffice Writer does.

Thus, in these situations as an author, you can turn to which of the spellings is in common use. I first looked at some of the statistics books on my shelves. I selected 14 (bio)statistics textbooks and checked the index and if present, chapters on graphics for term usage.

Table 4.5.1. Frequency of use of different terms for scatter plot in 14 (bio)statistics books currently on Mike’s shelves.

spelling	number of statistical texts	frequency
scatter diagram	2	0.144
scatter plot	5	0.357
scattergram	1	0.071
scatterplot	5	0.357
XY plot	0	0.071

Not much help; basically, it is a tie between “scatter plot” and “scatterplot.”

Next, I searched six journals for the interval 1990 – 2016 for use of these terms. Results are presented in Table 4.5.2, along with journal impact factor for 2014 and number of issues.

Table 4.5.2. Impact factor and number of issues 1990 – 2016 for six science journals.

Journal	Impact factor	Issues
The BMJ	17.445	1374
Ecology	5.175	271
J Exp Biol	2.897	540
Nature	41.456	1454
NEJM	55.873	1377
Science	33.611	1347

My methods? I used the journal’s online search functions for the various usages for scatter plot, and the results are shown in Figure 4.5.9.

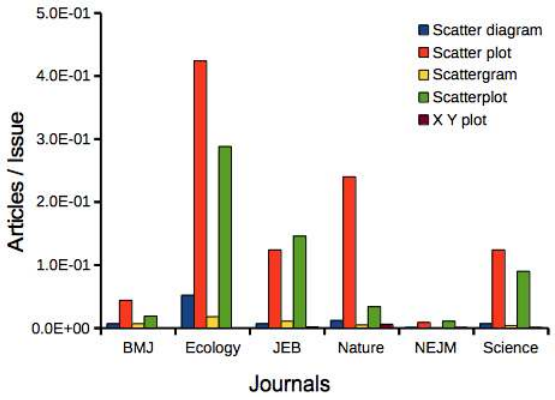


Figure 4.5.9: Usage of terms for X Y plots in research articles normalized to number of issues, in six journals between 1990 and 2016.

The journals have different numbers of articles; I partially corrected for this by calculating the ratio number of articles with one of the terms divided by the number of issues for the interval 1990 – 2016. It would have been better to count all of the articles, but even I found that to be an excessive effort given the point I'm trying to make here.

Not much help there, although we can see a trend favoring "scatter plot" over any of the other options.

And finally, to completely work over the issue I present results from use of Google's Ngram Viewer. Ngram Viewer allows you to search words in all of the texts that Google's folks have scanned into digital form. I searched on the terms in texts between 1950 and 2015, and results are displayed in Figure 4.5.10 and Figure 4.5.11.

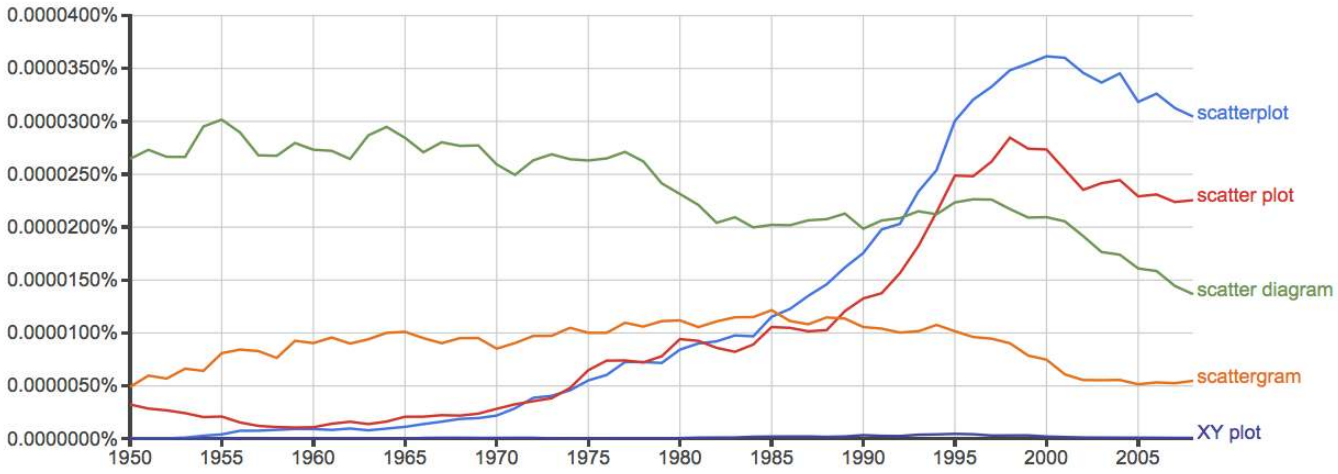


Figure 4.5.10: Results from Ngram Viewer for American English, "scatterplot" (blue), "scatter plot" (red), "scatter diagram" (green), "scattergram" (orange), and "XY plot" (purple).

And the same plot, but this time for British sources:

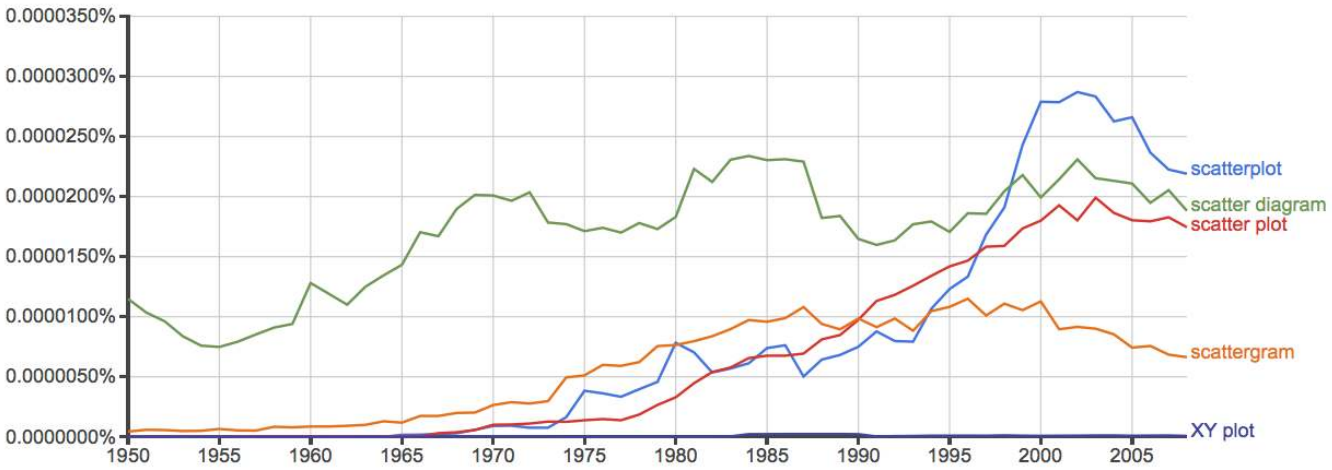


Figure 4.5.11: Results from Ngram Viewer for British English.

Conclusion? It looks like “scatterplot” (blue line) is the preferred usage, but it is close. Except for “scattergram” and “XY plot,” which, apparently, are rarely used. After all of this, it looks like you’re free to make your choice between “scatterplot” or “scatter plot.” I will continue to use “scatter plot.”

Questions

1. Using our Comet assay data set (Table 1, Chapter 4.2), create scatter plots to show associations between tail length, tail percent, and olive moment.
2. Explore different settings including size of points, amount of white area, and scale of the axes. Evaluate how these changes change the “story” told by the graph.

4.6: Adding a second Y axis

Introduction

Scatter plots are used to show association between two continuous variables. However, it is not uncommon to have a third variable for which association between the same X variable is expected. Thus, a common scatter plot type includes a second Y axis. These graphs are a bit more involved to make, regardless of which application used. The purpose of this short section is to provide a way to create a plot with two Y axes against a common X axis. Additional plotting options are also provided and explained.

The data set

As I write this note it is September 4, 2019, peak time for hurricanes. Hurricane Dorian passed the Bahamas as a category 4 storm heading for Florida. For context I include a screenshot of imagery from NOAA; you can see Dorian along the Florida coastline as well as four additional storms lined up from the coast of Africa across the Atlantic Ocean (Fig. 4.6.1).

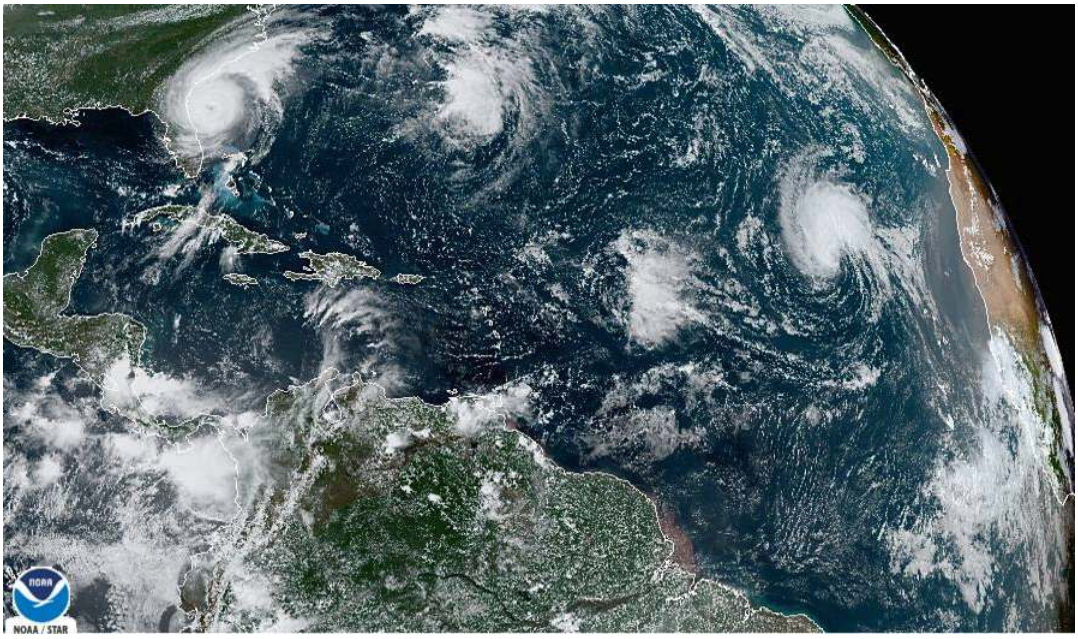


Figure 4.6.1: Screenshot from NOAA GOES-East – Sector view: Tropical Atlantic – GeoColor, 4 September 2019.

The data set here is simply the number of Atlantic Ocean Category 1, 2, 3, 4, or 5 storms since 1900, tabulated by decade. Storms are categorized by wind speed according to the Saffir-Simpson Hurricane Wind Scale, or SSHWS. Additional data of the average levels of carbon dioxide (CO_2), a greenhouse gas, measured at Mauna Loa and average global temperature index from NOAA were also acquired for the same period. (The temperature index is called anomaly data, as it is the difference of temperature between the average temperatures between 1950 and 1980).

```
#Create the time data, 12 decades, with 2010 based on 8 years only -- up to 2018
decade <- seq(1900,2010, by = 10)
#Saffir-Simpson Hurricane Wind Scale; complied from Wikipedia, checked against NOAA
cat01 <- c(2,0,1,3,0,9,4,11,18,11,27,46)
cat02 <- c(2,0,1,3,10,4,5,4,3,13,9,8)
cat03 <- c(7,2,3,1,4,10,9,5,4,12,11,7)
cat04 <- c(1,6,5,8,9,10,9,5,5,10,15,8)
cat05 <- c(0,0,2,6,0,2,4,3,3,2,9,4)
#land ocean anomaly temperature index, NOAA
tempIndex <- c(-0.317,-0.329,-0.241,-0.123,0.042,-0.048,-0.028,0.034,0.247,0.387,0.5
#mauna kea CO2, mean by decade, record starts 1958
co2 <- c(NA,NA,NA,NA,NA,316.0,320.3,330.9,345.5,360.5,378.6,399.0)
```

Combine all of the variables into a single data frame.

```
storms <- data.frame(decade,tempIndex,cat01,cat02,cat03,cat04,cat05,co2)
head(storms)
```

Output from R should look like

```
  decade tempIndex cat01 cat02 cat03 cat04 cat05 co2
1   1900    -0.317     2     2     7     1     0  NA
2   1910    -0.329     0     0     2     6     0  NA
3   1920    -0.241     1     1     3     5     2  NA
4   1930    -0.123     3     3     1     8     6  NA
5   1940     0.042     0    10     4     9     0  NA
6   1950    -0.048     9     4    10    10     2 316
```

Create a new variable with all categories of storms summed by decade.

```
#Sum all of the storms
allHur <- cat01+cat02+cat03+cat04+cat05
#Add new variable to the data frame
storms$allHur <- allHur
#Attach the data frame so you don't have to keep referring to the data frame when you
attach(storms)
```

Now, create a simple scatter plot of number of hurricanes by decade (Fig. 4.6.2).

```
plot(decade,allHur,pch=16)
```

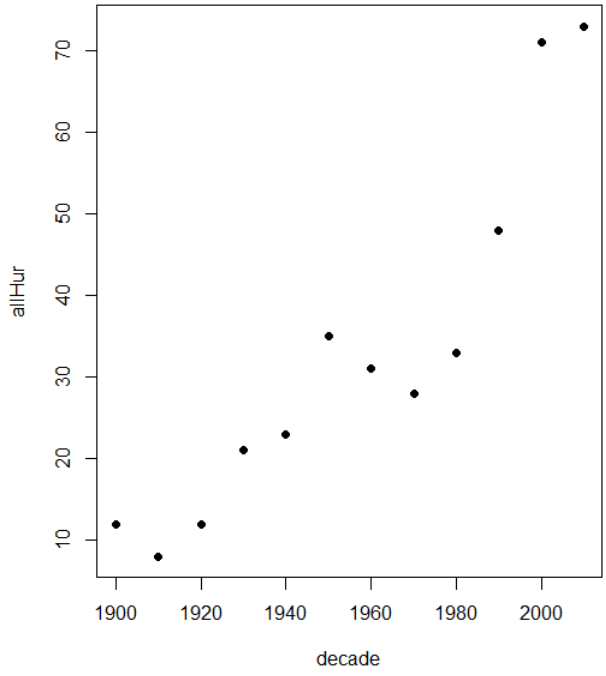


Figure 4.6.2: Plot of hurricanes from 1900 to present by decade.

A note on data management

This note is not needed for the plots. However, it's as good a place as any to comment about how to work with your data in R. For the hurricane data, I set the data frame `storms` as an unstacked (wide) worksheet. The code that follows shows how to use the `reshape2` package, part of the `tidyverse` set of packages for data manipulation, to convert the data frame from unstacked (wide) to stacked (long).

```
#code modified from examples presented at https://seananderson.ca/2013/10/19/reshape/
library(reshape2)
library(dplyr)
tryMe <- melt(storms, id.vars=c("decade", "tempIndex", "co2"), variable.name = "SSHWS
head(tryMe)
```

You can also take a stacked (long) worksheet and convert it to unstacked (wide) with `dcast()` .

```
tryAgain <- dcast(tryMe, decade + tempIndex + co2 ~ SSHWS)
```

Make the plots

At last, here's the code for adding a second Y axis. Code modified from https://www.r-bloggers.com/r-single-plot-with-two-different-y-axes/. The work flow begins by setting parameters of the plotting window, creating the first plot, followed up by adding the second plot (`par=TRUE`), setting graphing elements, then adding labels and a legend.

First plot example: Total number of hurricanes by decades, with Temperature Index by decades. Number of hurricanes represented on first (left) axis and Temperature Index represented on second (right) axis (Fig. 4.6.3).

```
par(mar = c(5,5,2,5))
#Create the first plot
plot(tryAgain$decade, tryAgain$allHur, pch=16, cex=1.5, col="black", xlab="Decades",
```

the syntax par(mar = c(bottom, left, top, right))

```
#Add the second plot
par(new=T)
plot(tryAgain$decade, tryAgain$tempIndex, type="l", lty="dashed", lwd=2, col="red3",
axis(side=4)
mtext(side=4,line=3,"Temperature Index")

#Add a legend
legend("topleft",legend=c("Hurricanes", "Temperature Index"),lty=c(0,2), pch=c(16,NA
```

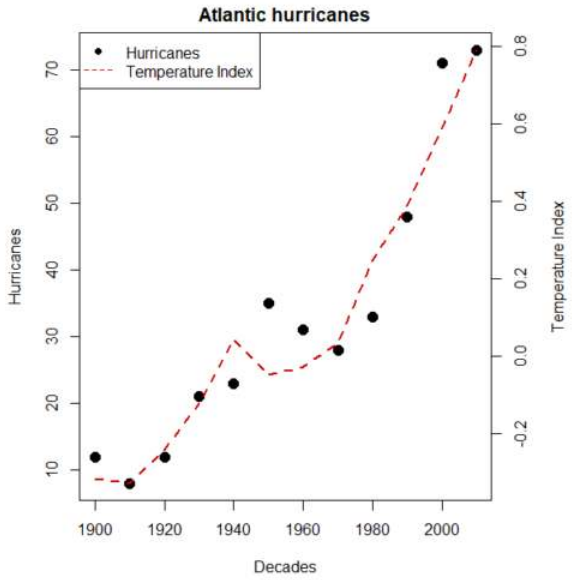


Figure 4.6.3: Total number of hurricanes by decades, with Temperature Index by decades. Number of hurricanes represented on first (left) axis and Temperature Index represented on second (right) axis.

Next plot example: Total number of hurricanes by decades, with Atmospheric CO_2 measured at Mauna Loa by decades. Number of hurricanes represented on first (left) axis and Atmospheric CO_2 represented on second (right) axis (Fig. 4.6.4).

```
par(mar = c(5,5,2,5))
plot(tryAgain$decade, tryAgain$allHur, pch=16, cex=1.5, col="black", xlab="Decades",
```

Create space for the second plot in the same frame (use T or TRUE, both work).

```
par(new=T)
```

Code for the second plot. Note that `axes=F` ("F" or "FALSE" works) is used to suppress printing axes — Recall that lines were already created by the first plot; if you do not add this, then additional lines are added to the plot."NA" is used to suppress adding labels; if you do not add this code, then labels will be printed over the existing labels created for the first plot. The code axis=4 sets the right-hand Y axis to active. The next code, `mtext()` is used to place the axis label for the second Y axis.

```
plot(tryAgain$decade, tryAgain$co2, type="l", lty="dashed", lwd=2, col="red3", axes=F
axis(side=4)
mtext(side=4,line=3,"Mean CO2, ppm")
```

Add the legend.

```
legend("topleft",legend=c("Hurricanes", "CO2"),lty=c(0,2), pch=c(16,NA), col=c("black
```

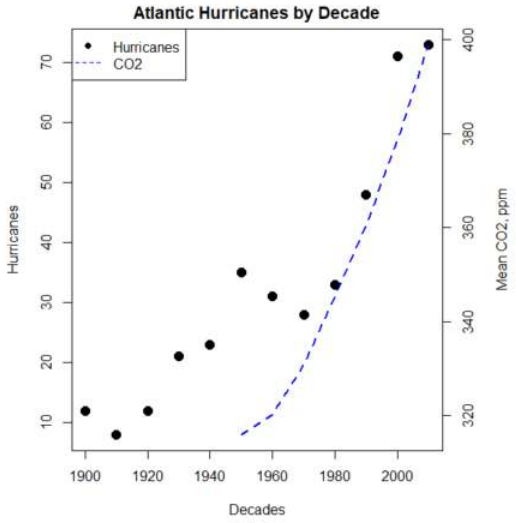


Figure 4.6.4: Total number of hurricanes by decades, with Atmospheric CO_2 measured at Mauna Kea by decades. Number of hurricanes represented on first (left) axis and Atmospheric CO_2 represented on second (right) axis.

Go ahead and try these plots. Change the settings (e.g., `lty` , `pch` , `cex` , `col`), and note how the graphs look.

Questions

1. Using the `plot()` command, make the following graphs:
   - One scatter plot for each category of storm by decade
   - Explore the kinds of graph elements available by changing `pch` values. Create your own
   - Change point size by changing valued for `cex`
2. Create a new plot with Decade on X axis, Temperature Index on first Y axis, and CO_2 on the second Y axis.
   - Include a legend for the plot.

4.7: Q-Q plot

Introduction

Use of graphs by a data analyst may serve different purposes: communication of results or as diagnostics. The Q-Q plot is one example of a graph used as a diagnostic.

The quantile-quantile, or Q-Q plot, is a probability plot used to graphically compare two probability distributions. In brief, a set of intervals for the quantiles is chosen for each sample. A point on the plot represents one of the quantiles from the second distribution (y value) against the same quantile from the first distribution (x value).

A common use of Q-Q plot would be to compare data from a sample against a normal distribution. If the sample distribution is similar to a normal distribution, the points in the Q–Q plot will approximately lie on the line $y = x$.

R code

In R, the Q-Q plot can be obtained directly in Rcmdr.

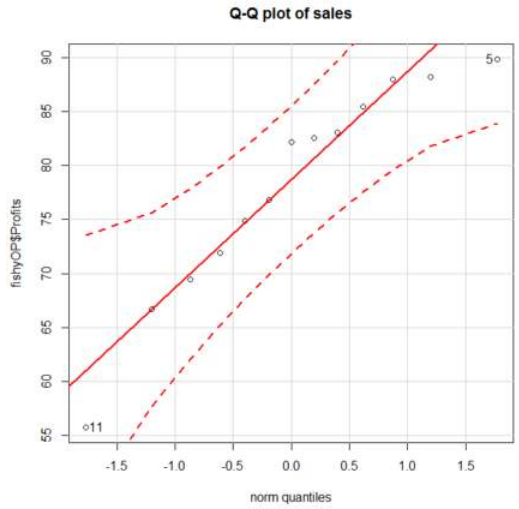


Figure 4.7.1: A Q-Q plot, the default command in Rcmdr.

Rcmdr: **Graphics → Quantile-comparison plot…**

After choosing the variable (in this case, Sales), click on Options tab and make additional selections before making the graph. Here, we selected normal distribution.

Figure 4.7.2: Screenshot of R Commander menu for Q-Q plot.

Another version is available in the `KMggplot2` package.

Questions

1. What is a Q-Q plot used for in statistics?
2. Looking at the plot in Figure 4.7.1, explain why the confidence lines get further and further away from the straight line.

4.8: Ternary plots

Introduction

Ternary plots, called **de Finetti diagram** in population genetics, is used to display three ratio variables that, together, sum to one. For example, display frequency of the three genotypes of a one gene, two allele system in a population.

Download the package `Ternary` from the R mirror. From the Ternary package, we can get a blank plot by simply calling the function `TernaryPlot()` . R returns the blank plot to the Graphics window (Fig. 4.8.1).

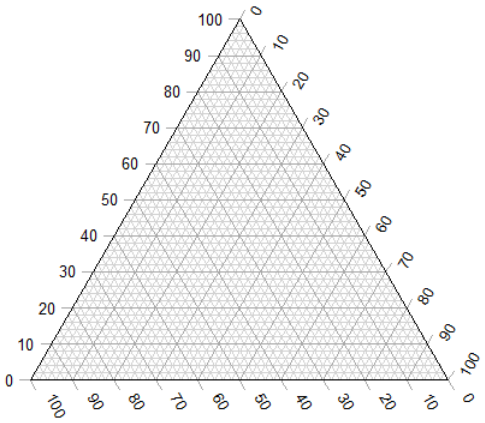


Figure 4.8.1: Blank Graphics window with initial ternary plot.

The basic ternary plot is shown in Figure 1. Running from one corner to another you can see how the frequencies range from 0 to 100%. While we can use the `Ternary` package, other packages allow you to make ternary plots too, including `HardyWeinberg.` This package includes several useful tests of the Hardy Weinberg model for population genetics data, so we'll use that package.

Or example will use the `HWTernaryPlot` function in the `HardyWeinberg` package. Before proceeding with the example, download and install the package.

> **Note:**
>
> A nice site on ternary plots in Microsoft Excel (24 steps!) is provided at chemostratigraphy.com. Instructions also worked for LibreOffice Calc (pers. obs.). Take a look at www.ternaryplot.com for a really nice online plot builder.

Example. Recall your basic population genetics, for a locus with 2 alleles with frequency p and q in the population, and given Hardy-Weinberg assumptions apply (e.g., no evolution!), then expected genotype frequencies are given by expanding $(p+q)^2 = 1$.

Consider a population genetics example using Skittles (Fig. 4.8.2).

Figure 4.8.2: A few Skittles® candies.

For several bags, count the greens (p) and the oranges (q). Data for 17 mini bags are reported in Table 4.8.1.

Table 4.8.1. Counts of green and orange Skittles from 17 mini bags.

Bag	GREEN	ORANGE
bag1	4	2
bag2	8	2
bag3	3	3
bag4	3	4
bag5	5	7
bag6	5	1
bag7	13	5
bag8	4	2
bag9	6	3
bag10	3	2
bag11	5	4
bag12	9	9
bag13	0	2
bag14	7	3
bag15	5	4
bag16	6	2
bag17	2	3

Next, we calculate the genotype frequencies from our counts. For example, for bag1, $p = 4/6$ and $q = 2/6$. We can imagine a diploids at the locus: GG, GO, and OO, with frequencies p^2, $2pq$, and q^2. The frequencies for the three genotypes are shown in Table 4.8.2.

Table 4.8.2. Genotype frequencies for our hypothetical population of Skittle diploid critters.

Bag	p^2	$2pq$	q^2
bag1	0.44	0.44	0.11
bag2	0.64	0.32	0.04
bag3	0.25	0.50	0.25
bag4	0.18	0.49	0.33
bag5	0.17	0.49	0.34
bag6	0.69	0.28	0.03
bag7	0.52	0.40	0.08
bag8	0.44	0.44	0.11
bag9	0.44	0.44	0.11
bag10	0.36	0.48	0.16
bag11	0.31	0.49	0.20
bag12	0.25	0.50	0.25
bag13	0.00	0.00	1.00

Bag	p^2	$2pq$	q^2
bag14	0.49	0.42	0.09
bag15	0.31	0.49	0.20
bag16	0.56	0.38	0.06
bag17	0.16	0.48	0.36

For the plot, the `HWTernaryPlot` function expects counts, not frequencies of three genotypes of a gene in a population, with genotype frequency that sums to one. Table 3 shows calculated genotype data, assuming 20 Skittle diploid critters per bag.

Table 4.8.3. Expected genotype counts.

Bag	GG	GO	OO
bag1	9	9	2
bag2	13	6	1
bag3	5	10	5
bag4	4	10	7
bag5	3	10	7
bag6	14	6	1
bag7	10	8	2
bag8	9	9	2
bag9	9	9	2
bag10	7	10	3
bag11	6	10	4
bag12	5	10	5
bag13	0	0	20
bag14	10	8	2
bag15	6	10	4
bag16	11	8	1
bag17	3	10	7

Example

Create an R data.frame called `skittles` from Table 4.8.3. Because `HWTernaryPlot` requires input only of the genotype data, remove the first column

```
dSkittles <- subset(skittles, select = -c(Bag) )
require(HardyWeinberg)
#Create a Ternaryplot
HWTernaryPlot(dLactose,100,pch=19, cex=2, region=1,hwcurve=TRUE, curvecols=c("red", "
```

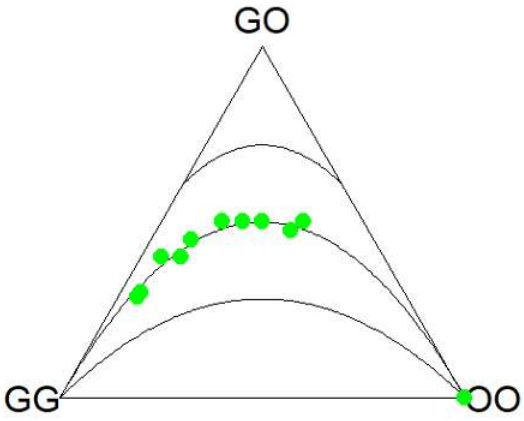


Figure 4.8.3: Ternary plot of our Skittle critter data.

What do we have? The function plots three convex curved lines. The green points are the heterozygote (GO) frequencies. They all fall on the middle line, as expected, because I had used HW to calculate frequency of the heterozygotes. If any population had numbers of observed heterozygotes different from expected values, then the population point would be represented by a red point and it would fall in one of the regions above or below the curved lines.

Question

1. Repeat the Skittles example, but replace with counts for purple (p) and red (q) candies (scroll down to data below, or click here).
    1. Optional: A more realistic example would be to draw 2 candies from Skittles bag and record the counts (e,g, how many purple+purple, purple+red, red+red pairs drawn), then make Ternary plots on the observed and not the expected frequencies.
2. Genetic example. The ternary plot is useful for displaying population genetic frequency data. For example, ability to digest lactose, i.e., lactase persistence, is in part due to genotype at SNP rs4988235 (Enattah et al 2002). Genotypes CC tend to be lactose intolerant, genotypes CT and TT are lactose tolerant. I gathered allele frequencies from the ALFRED database for several human populations, calculated genotype frequencies assuming Hardy-Weinberg equilibrium. I also created results for a hypothetical population "Madeup." (Scroll down to data below, or click here). Enter the data into R as a dataframe, e.g., `data.SNP` , copy R code from this page, make the necessary changes, and recreate the plot shown in Figure 4.8.4.
    - What do you conclude about the heterozygotes in Madeup population?

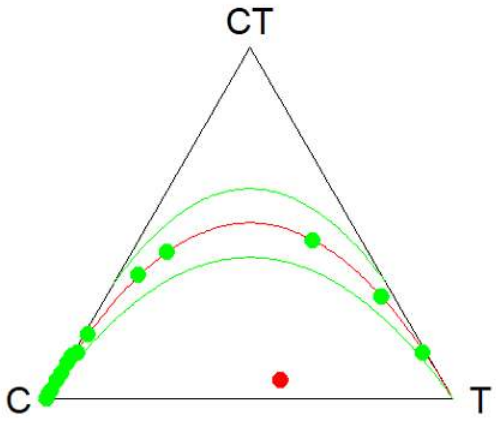


Figure 4.8.4: *rs4988235 genotype frequencies*, `data.SNP` .

3. Add a new row of data to your rs4988235 data set, `data.SNP` . CC= 4, CT = 10, TT = 6. The data were derived from frequencies reported in Figure 2, Baffour-Awuah et al 2016 (PMC4308731). To add a new row, modify the code below

```
data.SNP <- rbind(data.SNP, "PMC4308731" = c(4, 10, 6))
```

Create another ternary plot, and address whether or not this new data set shows heterozygotes in agreement with Hardy Weinberg assumptions.

Data sets

Skittles Data

Table 4.8.3. Counts of red and purple Skittles in 17 bags.

Bag	PURPLE	RED
bag1	3	3
bag2	8	4
bag3	2	4
bag4	2	1
bag5	7	6
bag6	3	2
bag7	5	5
bag8	2	5
bag9	4	4
bag10	3	5
bag11	3	2
bag12	8	6
bag13	2	5
bag14	7	6
bag15	3	1
bag16	2	4
bag17	4	1

SNP Data

Table 4.8.4. Genotype at SNP rs4988235 by population.

Population	C	CT	T
Mambila	99	1	0
Ben Amir	99	1	0
Zaramo	100	0	0
Bedouin	95	5	0
Druze	93	7	0
Kuwaiti	86	13	1
Dutch	12	45	43
Finns	3	29	68
Swedes	1	13	87
Bengali	87	13	1

Population	C	CT	T
Naga	100	0	0
Mala	88	12	0
Han	100	0	0
Japanese	100	0	0
Koreans	100	0	0
Cheyene	93	7	0
Pima	98	2	0
Maya	90	10	0
Brazilian	50	42	9
Chilean	60	35	5
Colombian	81	18	1
Madeup	40	5	55

4.9: Heat maps

Introduction

A **heat map** is a graph of data from a matrix (Wilksonson and Friendly 2009). Heat maps are common in many disciplines in biology, from ecology (e.g., diversity analyses) to genomics (e.g., gene expression profiling) to economics and demographics (Fig. 4.9.1). Instead of plotting numbers, color is used to communicate associations between cells in the rows and columns of the matrix.

Heat maps are useful for suggesting trends and typically do not require specialized knowledge to interpret. Provided that a color scale is defined, heat maps do a good job communicating trends. Viewers may rapidly make comparisons as they scan colors, from cold to hot.

Figure 4.9.1 provides a classic heat map: counties of the USA by percent ethnicity compared to "white" from Census.gov based on the 2010 census. The scale shows shades of blue, representing high percentages of white people (greater than 96.3%), to white, representing lower percentages of white people (less than 71%). Map is generated with mapping tool at United States Census Bureau TIGERweb.

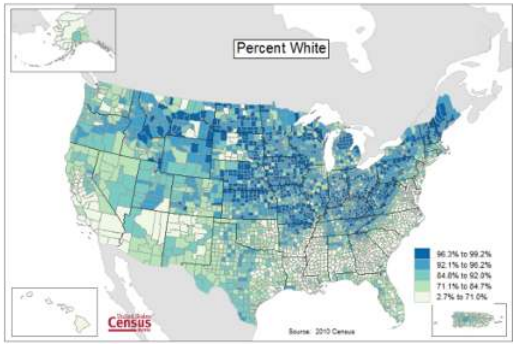


Figure 4.9.1: Heat map, 2010 USA population by county and percent ethnicity compared to white. Graph from census.gov.

Figure 4.9.2 shows gene expression results from a pilot study we did on metal exposure in cultured rat lung cells compared to cells without metal exposure (i.e., the control group). Genes were selected because of their role in the epithelial-mesenchyme transition, EMT. The color scale is typical for such studies: green represents down-regulation, red indicates up-regulation compared to the controls. Black is used to show no difference between treatment and control cells.

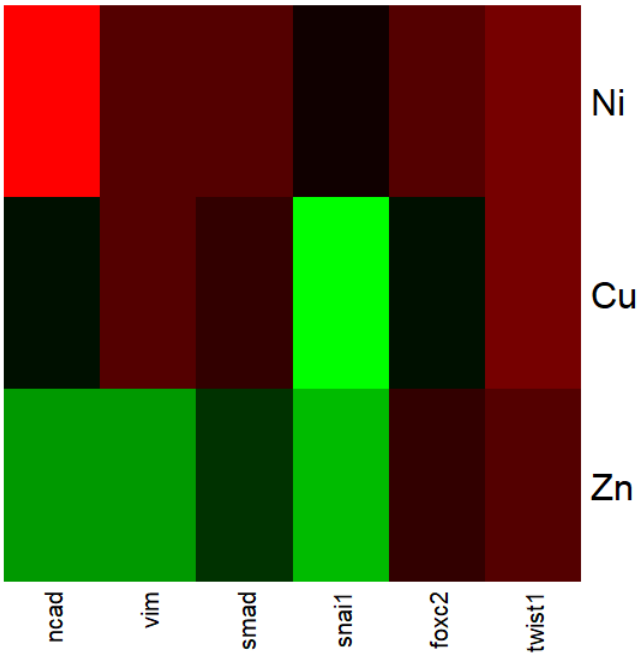


Figure 4.9.2: Heat map of gene expression in cultured rat lung cells exposed to metals.

Heat maps are good at directing the viewer to areas of strong association between variables, or in the case of comparisons, to draw strong inferences about the association. However, their chief limitations include gradations between colors; like pie charts, it is difficult to interpret the importance of slight changes in color, and the very use of heat map colors does not imply statistical significance (Chapter 8). Some color palettes are poor choices for viewers who may be colorblind. A good source about colors is available in the Graphs section of Cookbook for R.

R and heat maps

Lots of specialized packages will do cluster to heat map. Functions include `heatmap`, `heatmap2`, `heatmap.plus`, `NeatMap`. We'll step through how to make a heat map with another pilot study data from our lab.

heatmap(). Here's another heat map, percent DNA in tail from Alkaline Comet Assay (Figure 4.9.3). The same cultured cell line, a rat immortalized Type 2-like alveolar lung cell line L2 cells, were grown in media containing witch-hazel tea, a dilute copper solution, or both witch-hazel tea and copper (unpublished data). The hypothesis was that there would be greater DNA damage in cells exposed to copper compared to cells in hazel tea or a combination of copper and hazel tea. Witch hazel is reputed to have antioxidant properties (Pietta et al 1998). A random sample of 10 cells were sampled from each treatment (between 30 and 60 cells counted for each treatment). Within each treatment values were placed in ascending order, so "Cell 1" corresponds to the lowest value for a measured cell in each treatment.

```
#data arranged in unstacked worksheet
data <- as.matrix(hazelCuUnstack)

#check the import
head(data)

         Copper       Hazel HazelCopper]
[1,] 0.02404672 0.007185706 0.02663191
[2,] 0.06711479 0.027020958 0.03181153
[3,] 0.12196060 0.037725842 0.03743693
[4,] 0.13308991 0.044762867 0.03851548
[5,] 0.13344032 0.045809398 0.18787608
[6,] 0.17537831 0.060942269 0.19494708

#make the heat map
heatmap(data)
```

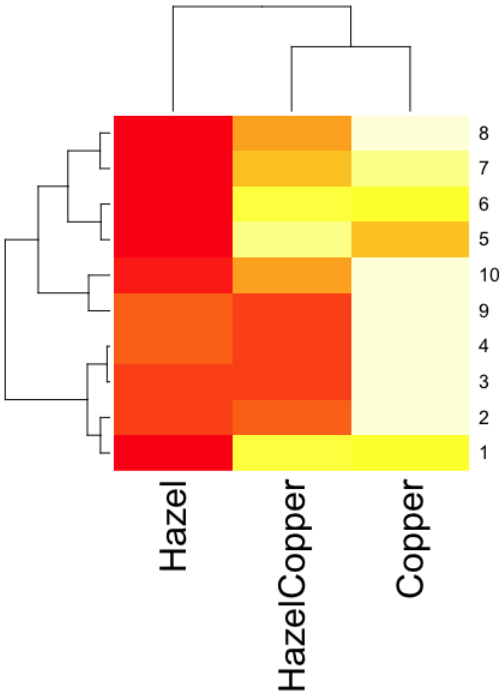


Figure 4.9.3: A simple heat map generated by `heatmap()` function, all default options.

The `heatmap()` function first runs a cluster analysis to group the cells by columns and rows — so similar cells are grouped together. The row and column **dendrograms** are default; your data are rearranged by the clustering procedure. To generate the heatmap without the dendrograms, add the following to the R code.

```
heatmap(data, Rowv = NA, Colv = NA)
```

ggplot2 and aes(). Not straight forward, but `ggplot2` (and therefore the Rcmdr plug-in `KMggplot2`) can be used. The `aes` function is part of the "**aesthetic mapping**" approach (Wickham 2010). The example below takes the same data and introduces use of a custom color palette, `brewer.pal` . Uses `geom_tile` , but `geom_raster` can also be used.

```
library(RColorBrewer)
#Explore the color profiles available at http://colorbrewer2.org/#type=sequential&scl
?brewer.pal
hm1.colors <- colorRampPalette(rev(brewer.pal(9, 'RdYlGn')), space='Lab')
#the data set
hazelCu <- read.table("hazelCu.txt", header=TRUE, sep="t", na.strings="NA", dec=".",
#Confirm the import
head(hazelCu)

Cell Treatment TailPerc
1 1 C 0.02404672
2 2 C 0.06711479
3 3 C 0.12196060
4 4 C 0.13308991
5 5 C 0.13344032
6 6 C 0.17537831

#convert cell number to factor.
hazelCu <- within(hazelCu, {
Cell <- as.factor(Cell)
})

ggplot(hazelCu,aes(x=Treatment,y=Cell,fill=TailPerc)) + geom_tile() + coord_equal() -
```

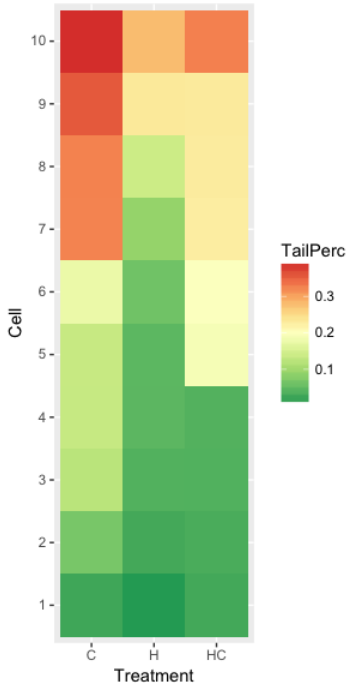


Figure 4.9.4: `ggplot()` and `aes()` functions used to generate a heat map. Colors from brewer.pal

The color scheme used in Figure 4.9.4 is common in gene expression studies: green is negative, cooler, while red is positive, hotter.

Questions

1. What are three advantages of heat map for communicating data.
2. What are three disadvantages of heat map for communicating data.
3. What color pallet is considered “color-friendly” for accessible visualization?

4.10: Graph software

Introduction

You may already have experience with use of spreadsheet programs to create bar charts and scatter plots. Microsoft Office Excel, Google Sheets, Numbers for Mac, and LibreOffice Calc are good at these kinds of graphs — although arguably, even the finished graphics from these products are not suitable for most journal publications.

For bar charts, pie charts, and scatter plots, if a spreadsheet app is your preference, go for it, at least for your statistics class. This choice will work for you; at least, it will meet the minimum requirements asked of you.

However, you will find spreadsheet apps are typically inadequate for generating the kinds of graphics one would use in even routine statistical analyses (e.g., box plots, dot plots, histograms, scatter plots with trend lines and confidence intervals, etc.). And, without considerable effort, most of the interesting graphics (e.g., box plots, heat maps, mosaic plots, ternary plots, violin plots), are impossible to make with spreadsheet programs.

At this point, you can probably discern that, while I'm not a fan of spreadsheet graphics, I'm also not a purist — you'll find spreadsheet graphics scattered throughout Mike's Biostatistics Book. Beyond my personal bias, I can make the positive case for switching from spreadsheet app to R for graphics is that the learning curve for making good graphs with Excel and other spreadsheet apps is as steep as learning how to make graphs in R (see Why do we use R Software?). In fact, for the common graphs, R and graphics packages like `lattice` or `ggplot2` make it easier to create publishing-quality graphics.

Alternatives to base R plot

This is a good point to discuss your choice of graphic software — I will show you how to generate simple graphs in R and R Commander which primarily rely on plotting functions available in the base R package. These will do for most of the homework. R provides many ways to produce elegant, publication-quality graphs. However, because of its power, R graphics requires lots of process iterations in order to get the graph just right. Thus, while R is our software of choice, other apps may be worth looking at for special graphics work.

My list emphasizes open source and or free software available both on Windows and macOS personal computers. Data set used for comparison from Veusz (Table 4.10.1).

Table 4.10.1. Observation counts in two categories.

Bees	Butterflies
15	13
18	4
16	5
17	7
14	2
14	16
13	18
15	14
14	7
14	19

1. GrapheR — R package that provides a basic GUI (Fig. 4.10.1) that relies on Tcl/Tk — like R Commander — that helps you generate good scatter plots, histograms, and bar charts. Box plot with confidence intervals of medians (Fig. 4.10.2).

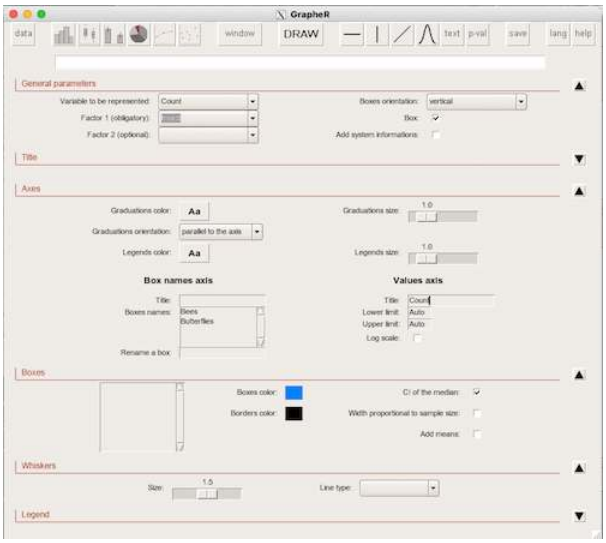


Figure 4.10.1: Screenshot of GrapheR GUI menu, box plot options

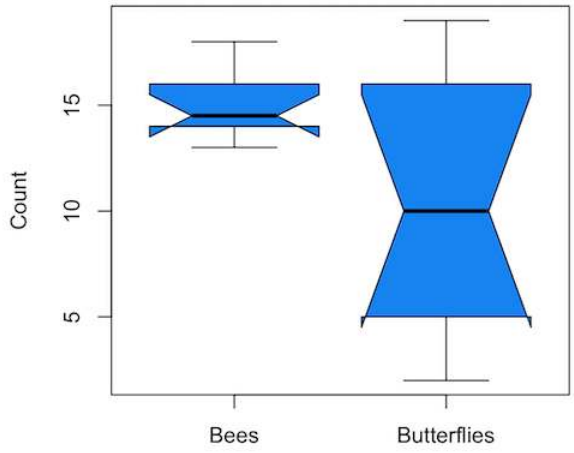


Figure 4.10.2: Box plot made with GrapheR.

2. RcmdrPlugin.KMggplot2 — a plugin for R Commander that provides extensive graph manipulation via the ggplot2 package, part of the Tidyverse environment (Fig. 4.10.3). Box plot with data point, `jitter` (Fig. 4.10.4)

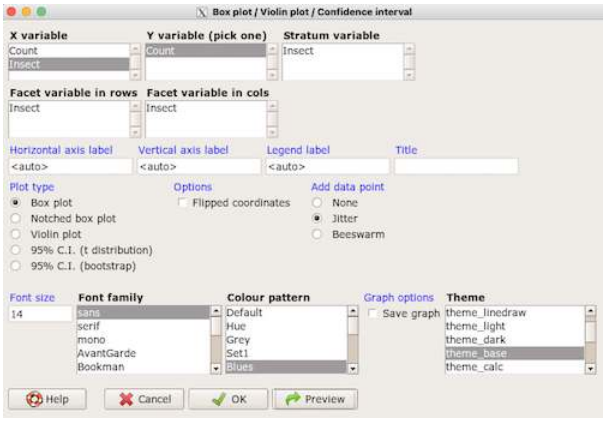


Figure 4.10.3: Screenshot of KMggplot2 GUI menu, box plot options

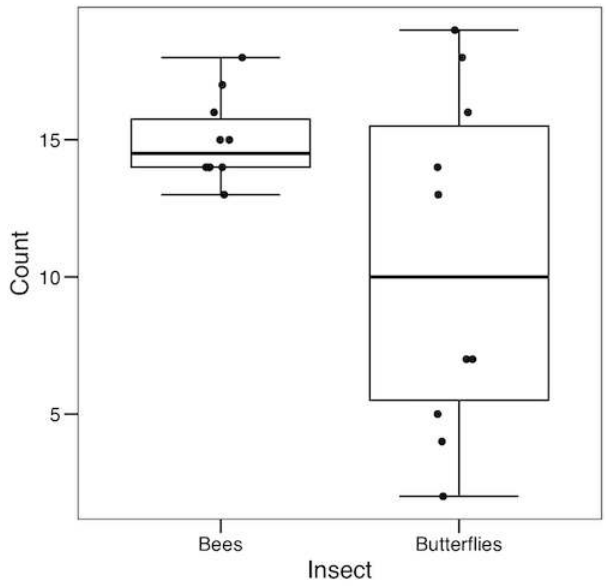


Figure 4.10.4: Box plot graph made with GrapheR with jitter applied to avoid overplotting of points.

Note:

If data points have the same value, overplotting will result — the two points will be represented as a single point on the plot. The **jitter** function adds noise to points with the same value so that they will be individually displayed. (Fig. 4) The **beeswarm** function provides an alternative to jitter (Fig. 5).

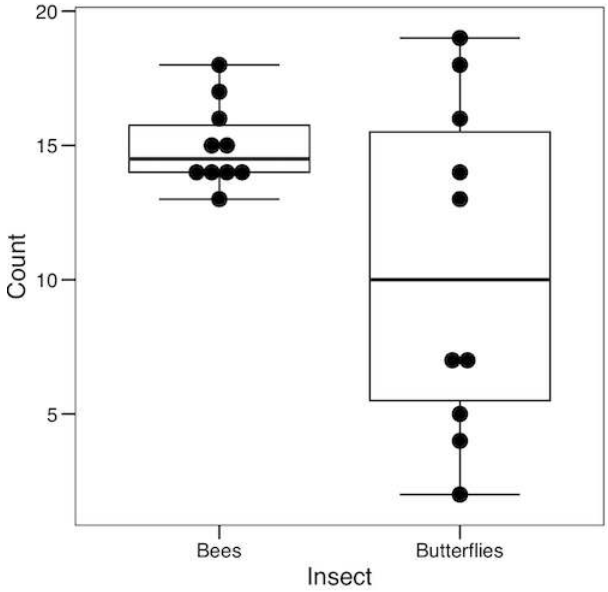


Figure 4.10.5: Box plot graph made with GrapheR with beeswarm applied to avoid overplotting of points.

3. A bit more work, but worth a look. Use plotly library to create interactive web application to display your data.

```
install.packages("plotly")
library(plotly)
fig <- plot_ly(y = Bees, type = "box", name="Bees")
fig <- fig %>% add_trace(y = Butterflies, name="Butterflies")
fig
```

code modified from example at https://plotly.com/r/box-plots/

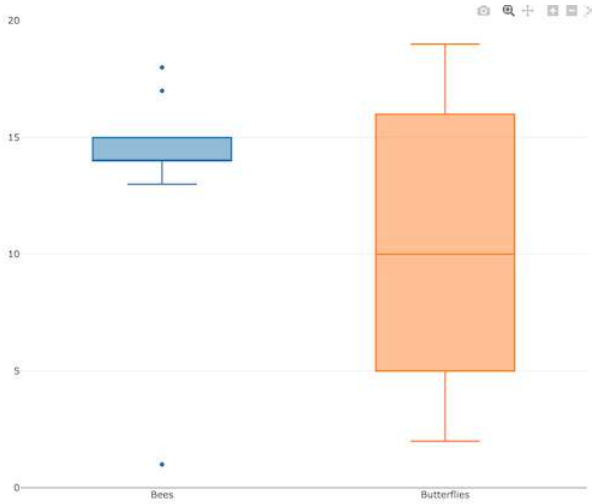


Figure 4.10.6: Screenshot of plotly box plot. Live version, data points visible when mouse pointer hovers.

4. Veusz, at https://veusz.github.io/. Includes a tutorial to get started. Mac users will need to download the dmg file with the `curl` command in the terminal app instead of via browser, as explained here.

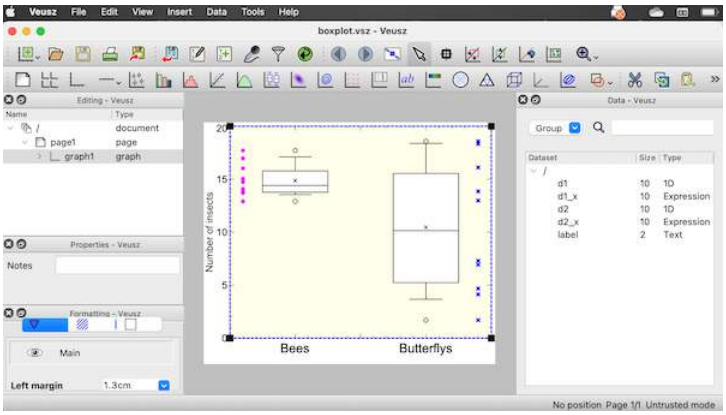


Figure 4.10.1: Copy and Paste Caption here. (Copyright; author via source)

5. SciDAVis is a package capable of generating lots of kinds of graphs along with curve fitting routines and other mathematical processing, https://scidavis.sourceforge.net/. SciDAVis is very similar to QtiPlot and OriginLab.

More sophisticated graphics can, and when you gain confidence in R, you'll find that there are many more sophisticated packages that you could add to R to make really impressive graphs. However, the point is to get the best graph, and there are many tools out there that can serve this end.

4.11: Chapter 4 References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17-21.

Chang, W. (2013). *R graphics cookbook*. O'Reilly Media, Sebastopol, CA.

Cumming, G. Error bars in experimental biology. *Journal of Cell Biology*, 177(1):7-11.

Cummings, P. (2003). Reporting statistical information in medical journals. *Archives of Pediatrics & Adolescent Medicine*, 157:321-323.

Drummond, G. B., & Vowler, S. L. (2011). Show the data, don't conceal them. *Advances in physiology education*, 35(2), 130-132.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., & Järvelä, I. (2002). Identification of a variant associated with adu .

Few, S. Are mosaic plots worthwhile? http://www.perceptualedge.com/articles/visual_business_intelligence/are_mosaic_plots_worthwhile.pdf

Fienberg. S. E. (1979). Graphical methods in statistics. *The American Statistician*, 33:165-178.

Friendly, M. (1994). Mosaic Displays for Multi-Way Contingency Tables. *Journal of the American Statistical Association* 89(425):190-200

Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 1886;15:246–263.

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach: turning tables into graphs. *The American Statistician*, 56(2), 121-130.

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in science education*, *47*(2), 183-210.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, *28*(1), 1–9. https://doi.org/10.18637/jss.v028.c01.

Klass, G. M. (2012). *Just Plain Data Analysis: Finding, presenting, and interpreting social science data*, 2nd ed. Rowman & Littlefield

Kumaravel, T. S., Vilhar, B., Faux, S. P., & Jha, A. N. (2009). Comet assay measurements: a perspective. *Cell biology and toxicology*, 25(1), 53-64.

Ovecka, G. D., Miller, G., Medeiros, D. M. Fatty acids of liver, cardiac and adipose tissues from copper-deficient rats. *Journal of Nutrition* 1988; 118:480-488

Pietta, P., Simonetti, P., & Mauri, P. (1998). Antioxidant Activity of Selected Medicinal Plants. *Journal of Agricultural and Food Chemistry*, 46(11), 4487–4490.

Scott (2009). Sturge's rule. *WIREs Computational Statistics* 1(3):303-306.

Streit, M., Gehlenborg, N. Points of View: Bar charts and box plots. *Nature Methods* 2014; 11(2):117.

Thring, T. S., Hili, P., & Naughton, D. P. (2011). Antioxidant and potential anti-inflammatory activity of extracts and formulations of white tea, rose, and witch hazel on primary human dermal fibroblast cells. *Journal of Inflammation*, 8(1), 27.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Chesire, Connecticut 06410.

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS biology*, 13(4), e1002128.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*, 2nd edition. Springer, New York, NY.

Wickham, H., Grolemund, G. (2016) *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Sebastopol, CA.

Wickham, H. (2010). A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics*, 19(1), 3–28.

Wilkinson, L., Friendly, M. (2009). The history of the cluster heat map. *American Statistician* 63:179-184.

CHAPTER OVERVIEW

5: Experimental Design

Introduction

During this course, you will learn about statistics, yes, but my hope and goal for your experience in this class is much more than that: **statistical reasoning**. To have statistical reasoning skills, you need to be comfortable with the context of how data are acquired, i.e., **data acquisition**. It's not just about knowing how an instrument performs, e.g., under or over the **limits of quantification** characteristics of the instrument lead to censored **missing values**, missing not at random (MNAR). In a broad stoke view, data are obtained from two kinds of studies: **observational studies** and **experimental** (manipulative) **studies**. what a correctly designed experiment can tell you about the world, and how a poorly designed experiment works against you. At the end of the semester, you should be familiar with the issues of Randomization, Control, Independence, and Replication.

The basics explained

Experimental design is a discipline within statistics concerned with the analysis and design of experiments. Design is intended to help research create experiments such that cause and effect can be established from tests of the hypothesis. We introduced elements of experimental design in Chapter 2.4. Here, we expand our discussion of experimental design.

We begin our discussion with review of definitions and an outline of an important concept in statistics. First, we need to be clear about why we measure or conduct experiments. We do so to collect data (datum is the singular) about a characteristic or trait from a population. Data are observations: they include observations and measurements from instruments. At their best, data are the "facts" of science.

Measurement, Observation, Variables, Values

Measurement is how we as scientists acquire our data (Chapter 3.4). The process of measuring involves the assignment of numbers, codes, or labels to observations according to rules established prior to any data collection (Stevens 1946, Houle et al 2011). **Observations** refer to the units of measurement, whereas **variables** are the characteristics or traits that are measured. A **value** refers to the particular number, score, or label assigned to a particular sample for a variable. Variables are generic, the thing being measured, values are specific to the subject or sample being measured for the variable. Each discipline in biology has its own set of variables and samples may or may not have different values for each variable measured. Variables are summarized as a statistic (e.g., the sample mean), which is a number taken to estimate a **parameter**, which pertains to the population. Variables and parameters in statistics were discussed in Chapter 3.4. Because numbers or scores or labels can be assigned according to different rules, this means that variables may be measured on different kinds of scales or data types. The different kinds of data types were presented in Chapter 3.1.

Missing values

A missing value refers to lack of a value for an observation or variable. Missing values can affect analysis and many R algorithms are sensitive or may fail to run in the presence of missing values. **Censored values** include observations for which only partial information is available. Missing data may be of three kinds, and one of them, **missing not at random** (MNAR), can influence the analysis. MNAR implies some observations are missing because of a **systematic bias**. Instrument **limits of quantification** are an example of systematic bias — for example, spectrophotometric absorbance readings of zero for a colorimetric assay (e.g., Bradford protein assay) may not represent complete absence of the target, but rather, the lower detection limits of the instrument or the assay method (0.1 mg protein/mL in this case).

The other kinds of missing values are **missing completely at random** (MCAR) and **missing at random** (MAR). MCAR implies there is no association between any element of the experiment and the absence of a value. Analysis on MCAR data sets may support unbiased conclusions. MAR includes the random errors that occur during data acquisition: date may be lost because of operator error. Analysis of MAR data sets, as with MCAR data sets, may still result in unbiased conclusions; obviously, the size of the data set influences whether this claim holds. In some cases, missing values can be replaced by **imputed values**.

Cause & Effect

Observations or measurements gathered under controlled conditions (experiments) are essential if we are to answer questions about populations, to separate **cause and effect**, where one or more events is directly the result of another factor, as opposed to **anecdote**, a story about an event which, by itself, cannot be used to distinguish cause from association (e.g., spurious correlations, see Ch 16.2). Recall that anecdotal evidence typically comes from personal experience, where observations may be obtained by non-systematic methods. Well-designed experiments, in the classical sense, permit discrimination among competing hypotheses in large part because observations are collected according to strict rules. Well-developed hypotheses tested by well-designed experiments permit ruling out alternative explanations (sensu Pratt [1964] **strong inference**). Observational studies, or epidemiology studies if we are talking about investigations of risk assessment, also may contribute to discussions of cause and effect (see history of smoking by Doll 1998).

Medicine is replete with stories about how a patient showed a particular set of symptoms, and how a physician applied a set of diagnostic protocols. An **outcome** was achieved, and the physician reports the outcome and circumstances related to the patient to her colleagues. This is an example of a **case study**, and the focus of investigation is the individual, the patient. The doctor's report will sound like, "I tried the standard treatment given the diagnosis, and the patient's symptoms diminished, but later returned. I tried a higher dose, but the patient's symptoms persisted unabated." No inferences are made to a wider set of individuals and the report is anecdotal. If observations are made on several patients, this may be a **case series**.

In ecology, a field biologist may notice a six-legged adult frog (Scott 1999, Alvarez et al 2021) — since frogs typically have four legs, the six-legged frog attracts the biologist's eye and he jots down the circumstances in which the frog were found: relative humidity, air temperature, ground temperature, where the frog was found (on lower leaf of a philodendron plant). A water source near where the frog was found is tested for pH with a meter the biologist carries, and a water sample is taken for later testing of herbicides. The frog is collected so that it can be checked for skin parasites. Upon further inspection, the frog did indeed have parasites known to cause deformities in other frog species. However, note that this example too, is a case study. Although the biologist makes additional observations, any conclusions about why the frog has six legs is anecdotal.

From these case studies, no conclusions can be drawn. We cannot say why the patient failed to respond to treatment, nor can we say why the frog has six legs. Why? Because these are singular events, and a variety of explanations can be given as to their causes — importantly, no controls are available, so there's no way to distinguish among possibilities.

From such anecdotes, however, experiments can be designed. The physician may decide to recruit additional patients with the diagnosed illness and apply the standard treatment to see if her anecdote is a single, unique event, or more indicative of a problem with the treatment. The biologist may collect other frogs from the area near where he found the six-legged frog and check to see if they, too, have the parasites. If additional patients fail to respond to the treatment, then the singular even is more likely to be a phenomenon. If the normal frogs also have similar levels of the parasite then it is unlikely that these parasites caused the malformed frog. With this simple step (recruiting similar patients, finding additional frogs), we can begin to make inferences about cause and effect and in some cases, to generalize our findings.

This is the objective of most statistical procedures, the concept of sampling from a reference population and making distinctions between groups within the sample. The difference between observational and experimental studies then is how the subjects are selected with respect to the groups. In an experiment, the researcher controls and decides which subject receives the treatment; therefore, allocation to groups is manipulated by the researcher. In contrast, subjects included in an observational study have already been "assigned" to a group, but not by us. Assignment to groups such as smokers or non smokers, Type II diabetes or no diabetes, etc. is done by nature.

Now, I do not wish to imply that research that cannot be generalized back to a reference population are worthless. Far from it. In fact, there is a strong argument for specificity. For example, much basic biological research depends on work in model organisms, which in turn may be further partitioned into specific genetic lines (cf. discussion in Rothman et al 2013). And my goodness, what we have learned about the devastation to oceanic islands like Guam when the brown tree snake was introduced (Fritts and Rodda 1998). Strictly speaking, what has happened to Guam is a case history. But no one would argue that what has happened to Guam cannot happen to Hawaii and other oceanic islands (e.g., United States Federal law 384-108 "Brown Tree Snake Control and Eradication Act of 2004"). In other words, even from case histories, generalizations can sometimes be made.

There can also be real reasons to ignore the issue of generality. One benefit of specificity is experimental control. Transgenetic lines may differ by single gene knockout or by gene duplication, and clearly the aim of such studies is to evaluate the function (hence purpose) of the gene product (or its absence) on some phenotype. In this sense it may not seem important that a transgenic

mouse is not representative of a wild outbred mouse population. However, this argument is fundamentally one of expedience — such studies do result in specific results, results that cannot be generalized beyond the strains involved. It ignores the issue of genetic background — all of the genes that affect a trait in addition to the candidate gene under study (Sigmund 2000; Lariviere et al 2001). Transgenic mice of different inbred strains or their hybrids may have very different alleles at other genes that may influence a phenotype; hence, the results of the gene knockdown or other engineering result in different outcomes. Results of genetic manipulations on inbred strains, no matter how sophisticated, mean that the conclusions are strain- or hybrid cross-specific. Thus, although technically and financially difficult, conclusions are better, more **generalizable**, when conducted with many different inbred lines and verified in outbred mouse populations precisely because genetic background often influences function of single genes (Sigmund 2000; Lariviere et al 2001).

"Causal criteria:" Logic of causation in medicine

This section is in progress. Just a list of key points and references

Throughout this text, emphasis on the power of experimentation is emphasized. Well-designed experiments should consider …

Henle-Koch's postulates (1877, 1882), developed from working on tuberculosis to report a set of causal criteria to establish link between a microorganism and a disease, are the following:

- The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.
- The microorganism must be isolated from a diseased organism and grown in pure culture.
- The cultured microorganism should cause disease when introduced into a healthy organism.
- The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

Robert Koch wrote these more than 100 years ago, so, clearly, understanding of infectious disease has improved. **Evan's postulates**, quoted from *A Dictionary of Epidemiology*, 5th edition (pp. 86-87):

1. Prevalence of the disease should be significantly higher in those exposed to the hypothesized cause than in controls not so exposed.
2. Exposure to the hypothesized cause should be more frequent among those with the disease than in controls without the disease —when all other risk factors are held constant.
3. Incidence of the disease should be significantly higher in those exposed to the hypothesized cause than in those not so exposed, as shown by prospective studies.
4. The disease should follow exposure to the hypothesized causative agent with a normal or log-normal distribution of incubation periods.
5. A spectrum of host responses should follow exposure to the hypothesized agent along a logical biological gradient from mild to severe.
6. A measurable host response following exposure to the hypothesized cause should have a high probability of appearing in those lacking this before exposure (e.g., antibody, cancer cells) or should increase in magnitude if present before exposure. This response pattern should occur infrequently in persons not so exposed.
7. Experimental reproduction of the disease should occur more frequently in animals or humans appropriately exposed to the hypothesized cause than in those not so exposed; this exposure may be deliberate in volunteers, experimentally induced in the laboratory, or may represent a regulation of natural exposure.
8. Elimination or modification of the hypothesized cause should decrease the incidence of the disease (e.g., attenuation of a virus, removal of tar from cigarettes).
9. Prevention or modification of the host's response on exposure to the hypothesized cause should decrease or eliminate the disease (e.g., immunization, drugs to lower cholesterol, specific lymphocyte transfer factor in cancer).
10. All of the relationships and findings should make biological and epidemiological sense.

Fredericks, D. N., & Relman, D. A. (1996). Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clinical microbiology reviews*, *9*(1), 18-33.

Correlation (association) does not imply causation, a well-worn truism in any application of critical thinking skills.

Note:

Association is the more general term for a possible relationship between two or more variables. A correlation in statistics generally refers to a linear association (Chapter 16); the aforementioned truism should be restated as association does not imply causation.

However, sometimes association does point to a cause. A familiar example is association between tobacco cigarette smoking causes lung cancer. Surgeon General Luthar Terry's 1964 report (link to document in National Library of Medicine) presented a strong case linking smoking to elevated risk of lung cancer and coronary artery disease.

Bradford Hill's guidelines to evaluate causal effects based on epidemiology (Hill 1965, see also Sussar 1999, Fedak et al 2015). They form a set of necessary and sufficient conditions, based on inductive reasoning.

1. Strength of association
2. Consistency of observed association
3. Specificity of association
4. Temporal relationship of the association
5. Biological gradient, e.g., a dose-response curve
6. Biological plausibility
7. Coherence, the cause and effect inference should not conflict with what is known about the etiology of a disease.

Follows and extends David Hume's (1739) causation criteria: association (Hill #1), cause precedes effect (Hill #4), direction of connection.

Validity

An obvious objective of research is to reach valid conclusions about fundamental questions. A helpful distinction between the specific and the generalizable experiment is to recognize there are two forms of validity in research: **internal validity** and **external validity** (Elwood 2013). Internal validity is the quality of a designed study that determines whether cause and effect can be determined. Random assignment of subjects to treatment groups enhances the internal validity of the study. External validity relates to how general the assessment of cause and effect can be to other populations. Thus, random sampling from a reference population has to do with whether or not the study has external validity.

Additional definitions

We proceed now with definitions. We use the term **population** in a special and restrictive way in statistics. Our definition includes the one you are already familiar with, but it also means more than that.

Populations are the entire group of individuals that you want to investigate. In statistics, the entire groups is actually the entire class with the observation — so if we are referring to the average body weight of house mice, we're actually referring to the body weight as the population — it's a subtle distinction, not essential for our introduction to biostatistics. When we conduct experiments and apply statistical tests on collected data, we generally intend to make inferences (draw conclusions) from our results back to the population.

Population has a strict application in statistics, but the definition also includes our general understanding of the word population. For example, examples of a population in the general sense that one may refer to include:

- the entire human population existing today.
- the entire collection of U.S. citizens.
- all the individuals in an entire species.
- all individuals in a population of a species (e.g., house mice in a dairy barn in Hawai'i).
- all of us in this class room (if we are *only* interested in *us*).

If you could measure the entire population then there would be no need to do (or learn) statistics! Populations usually are in the thousands, millions, or billions of individuals. Here, population is used in the everyday sense that we think of — a collection of individuals that share a characteristic.

A more formal definition of "population" in statistics reads as follows: A statistical population is the complete set of possible measurements on a trait or characteristic corresponding to the entire collection of sampling units for which conclusions are intended.

To conclude, in this class, when we talk about population, we will generally be using it in the everyday sense of the word. However, keep in mind that the definition is more restrictive than that and the key is to identify what sampling units are measured.

Conclusions

This is only the beginning, the basics of experimental design. Entire books are written on the subject, as you can well imagine. We will also return to the subject of Experimental Design throughout the book. We will return to random sampling in Chapter 5.5. Next we discuss distinctions between experiments and observational studies with respect to sampling of populations.

A bit of a disclaimer here before proceeding; while I cite several papers for examples in experimental design in Chapter 5, readers should not read into this that I am either criticizing or endorsing the published experiments. Experimental design will always have elements of compromise — the trick, of course, is knowing which choices influence validity (Thompson and Panacek 2006).

Questions

1. Define in your own words the following terms:
    - reference population
    - subjects
    - specific versus general conclusions
    - random sampling
    - convenience sampling
    - haphazard sampling
    - research validity
2. Revisit our cell experiment, “What is the sampling unit in the following cell experiment?” How would you change this experiment so that there will be biological and not just technical replication?
3. Describe the type of sampling for each research scenario described:
    - All African snails on a staircase at Chaminade University are collected on a Thursday evening.
    - All African snails on a staircase at Chaminade University are collected every Thursday evening for six months.
    - All African snails on all staircases at Chaminade University are collected.
    - African snails are studied in the lab, then returned to the areas from which they were collected. Days later, the researcher collects snails from the same area.
    - African snails are studied in the lab, then returned to the areas from which they were collected. Days later, the researcher collects snails from a different area.
    - The Chaminade University campus is divided into grids. Grids that include stairwells are marked. Before collecting snails, the researcher randomly selects from the list of grids and searches for snails only in those grids selected from the list.

Figure 5.1: Giant African Snail (*Lissachatina fulica*, formerly *Achatina fulica*). Image by M. Dohm.

4. A researcher wishes to study the effects of salt on mosquito larval survival. He works with *Aedes* species, mosquitos that are characterized as “container-breeding” – their larvae develop where water accumulates in tree holes or indentations in rock, or even in the containers left by humans (e.g., tires, flower vases or planters). His preliminary experiment is outlined in the following table. The last column indicates the measurement that he plans to record. Identify the sampling unit. Identify the experimental unit

5. Consider the Hermon Bumpus House sparrow survival data set (described at Field Museum (Chicago, IL) and American Ornithology Society), famous as an early example of natural selection. A storm on 1 February 1898, in Providence Rhode Island left dozens of house sparrows on the ground. Birds were collected and brought to Bumpus's, 136 in all. Seventy two revived, 64 died. Bumpus identified the sex and measured nine morphological traits of each bird. Bumpus concluded from his graphs that males survived better than females and that shorter, lighter birds with longer legs, wings and sternums and larger brain size ("skull width") also survived better. Which type of study is the Bumpus study? Select one:
    - Case study
    - Anecdote study
    - Case control study
    - Cohort study
    - Cross-sectional study
6. This next scenario may be evaluated by you for potential sources of bias. Review the list of bias listed above. A researcher wants to do a population count of feral cats on campus. Feral cats are active at night, so he decides to set up a feeding station near a light post. The researcher sits all night in a parked car yards and watches the feeding station for visits by cats. The researcher repeats these observations over the course of a week, moving the feeding station to different campus locations each night, and reports the total number of cats seen during the week as an estimate of the population size. Be able to discuss this study in terms of potential and actual bias.

5.1: Experiments

Introduction

With the background behind us, we outline a designed experiment. Readers may wish to review concepts presented previously, in

- Chapter 2.4: Experimental design and rise of statistics in medical research.
- Chapter 3.4 – Estimating parameters
- Chapter 3.5 – Statistics of error

Experiments have the following elements:

1. Define a **reference population** (e.g., all patients with similar symptoms; all frogs in the population from which the malformed frog belonged.
2. Define sample unit from the reference population (e.g., as many patients as can be screened by the physician are recruited; all frogs visible to the biologist are captured).
3. Design sampling scheme from population (e.g., **random sampling** vs. **convenience/haphazard sampling**).
4. Agree on the primary outcome or endpoint to be measured and whether additional secondary outcomes are to be observed and collected.
5. Separate the sample into groups so that comparisons can be made (e.g., the illness example doesn't exactly follow this scheme but rather, all are given the same treatment and responses are followed; a more typical example from biomedicine would be **random assignment** of individuals to one of two groups — half will receive a placebo, half will receive drug A; for the frog example, this analogy fits well — two groups — normal frogs and the the single abnormal frog),
6. Devise a way to exclude or distinguish from the possible explanations, or **alternate hypotheses** e.g., for the physician, she will record how many of her patients fail to respond to treatment; for the biologist, noting presence or absence of parasites between the normal and abnormal frog defines the groups).

A basic experimental design looks like this.

1. One treatment, with two levels (e.g., a control group and an experimental group)
2. A collection of individuals recruited from a defined population, esp. by random sampling.
3. Random assignment of individuals to one of the two treatment levels.
4. The treatments are applied to each individual in the study.
5. A measure of response (the **primary outcome**) and additional features (**secondary outcomes**) are recorded for each individual.

Contrast with design of observational studies

Note, importantly, that in **observational studies** no matter how sophisticated the equipment used to measure the outcome variable(s), key steps outlined above are missing. Researchers conducting observational studies do not control allocation of subjects to treatment groups (steps 4 and 3 in the two above lists, respectively). Instead, they may use **case-control** approaches, where individuals with (case) and without (control) the condition (e.g., lung cancer) are compared with respect to some candidate risk factor (e.g., smoking). Both case and control groups likely have members who smoke, but if there is an association between smoking and lung cancer, then more smokers will be in the case group compared to the control group.

A **cohort study**, also a form of observational study, is similar to case-control except that the outcome status is not known. A cohort study includes, in our example, smokers and non smokers who share other characteristics: age, medical history, etc. The other kind of design you will see is the **cross-sectional** study. Cross-sectional studies are descriptive studies, and, therefore also observational. **Primary outcomes**, along with additional characteristics and outcomes, are measured for a representative subsample, or perhaps even an entire population. **Cross-sectional studies** are used to absolute and relative risk rates. In ecology and evolutionary biology, cross-sectional studies are common, e.g., comparisons of species for metabolic rate (Darveau et al 2005) or life history traits (Jennings et al 1998), and specialized statistical approaches that incorporate phylogenetic information about the species are now the hallmark of these kinds of studies.

A word on outcomes of an experiment. Experiments should be designed to address an important question. The outcomes the researcher measures should be directly related to the important question. Thus, in the design of clinical trials, researchers distinguish between primary and secondary outcomes or endpoints. In educational research the primary question is whether or not students exposed to different teaching styles (e.g., lecture-style or active-learning approaches) score higher on an knowledge-

content exams. The primary outcome would be the scores on the exams; many possible secondary outcomes might be collected, including students' attitudes towards the subject or perceptions on how much they have learned.

Example of an experiment

We'll work through a familiar example. The researcher is given the task to design a study to test the efficacy in reducing tension headache symptoms by a new pain reliever. There are many possible outcomes: blurry vision, duration, frequency, nausea, need for and response to pain medication, and level of severity (Mayo Clinic). The drug is packaged in a capsule and a placebo is designed that contains all ingredients except the new drug. Forty subjects with headaches are randomly selected from a population of headache sufferers. All forty subjects sign the consent form and agree to be part of the study. The subjects also are informed that while they are participating in the research study on a new pain reliever, each subject has a 50-50 chance of receiving a placebo and not the new drug. The researcher then randomly assigns twenty of the subjects to receive the drug treatment and twenty to receive the placebo, places either a placebo pill of the treatment pill into a numbered envelope and gives the envelopes to a research partner. The partner then gives the envelopes to the patients. Both patients and the research partner are kept ignorant of the assignment to treatment.

We can summarize this most basic experimental design in a table, Table 5.1.1.

Table 5.1.1. Simple formulation of a 2×2 experiment *(aka 2×2 contingency table)*.

		Did the subject get better?		
		Yes	No	Row totals
Did the subject receive the treatment?	Yes	**a**	**c**	a + c
	No	**b**	**d**	b + d
	Column totals	a + b	c + d	N

where N is the total number, **a** is number of subjects who DID receive the treatment AND got better, **b** is number of subjects who DID NOT receive the treatment and DID get better, **c** is number of subjects who DID receive the treatment and DID NOT get better, and **d** is number of subjects who DID NOT receive the treatment and DID NOT get better.

And our basic expectation is that we are testing whether or not treatment levels were associated with subjects "getting better" as measured on some scale.

One possible result of the experiment, although unlikely, all of the negative outcomes are found in the group that did not receive the treatment (Table 5.1.2).

Table 5.1.2. One possible outcome of our 2×2 experiment.

Subject received the treatment	Subject improved	Subject did not improve
Yes	20	0
No	0	20

Results of an experiment probably won't be as clear as in Table 5.1.2. Treatments may be effective, but not everyone benefits. Thus, results like Table 5.1.3 may be more typical.

Table 5.1.3. A more likely outcome of our 2×2 experiment.

Subject received the treatment	Subject improved	Subject did not improve
Yes	7	13
No	2	18

Questions

1. At the time I am updating this page we are starting our fifth month of the coronavirus pandemic of 2019-2020. Daily it seems, we are also hearing news about “promising coronavirus treatments,” but as of this date, no study has been published that meets our considerations for a proper experiment. However, on May 1, the FDA issued an emergency use authorization for use of the antiviral drug remdesivir (Gilead), based on early results of clinical trials reported 29 April in *The Lancet* (Wang et al 2020). Briefly, their study included 158 treated with does of the antiviral drug and 79 provided with a placebo control. Both groups were otherwise treated the same. Improvement over 28 days was recorded: 62 improved with Placebo and 133 improved receiving does of remdesivir.
    1. Using the background described on this page, list the information needed to design a proper experiment. Using your list, review the work described in *The Lancet* article and check for evidence that the trial meets these requirements.
    2. Create a 2×2 table with the described results from *The Lancet* article.
2. Consider our hazel tea and copper solution experiment described in Chapter 3. The outcome variables (Tail length, tail percentage, Olive moment) are quantitative, not categorical. Create a table to display the experimental design.
3. For the migraine example, identify elements of the design that conform to the randomized control trial design described in Chapter 2.4.

5.2: Experimental units and sampling units

Introduction

Sampling units refer to the measured items, the focus of data collection; samples are selected from populations. Often, sampling units are the same thing as individuals. For example, if we are interested in the knowing whether men are more obese than women in Hawai'i, we would select individuals from the population; we would measure individuals. Thus, the sampling unit would be individuals, and the measurement unit would be percent body fat recorded for each individual. The data set would be the collection of all body fat measures for all individuals in the study, and we would then make inferences (draw conclusions) about the differences, if any, between adult males and females for body fat.

But sampling units can also refer to something more restrictive than the individual. For example, we may be interested in how stable, or consistent, a person's percent body fat is over time. If we take a body fat measure once per year over a decade on the same group of adults, then the sampling unit refers to each observation of body fat recorded (once per year, ten times for an individual), and the population we are most likely to be referring to is the collection of all such readings (ten is arbitrary — we could have potentially measured the same individual thousands of times).

In some cases the researcher may wish to compare groups of individuals and not the individuals themselves. For example, a 2001 study sought to see if family structure influenced the metabolic (glycemic) control of children with diabetes (Thompson et al 2001). The researchers compared how well metabolic control was achieved in children of single parents and two-parent families. Thus, the sampling units would be families and not the individual children.

Experimental units

Experimental units refers to the level at which **treatments** are independently applied in a study. Often, but not always, treatments are applied directly to individuals and therefore the sampling units and experimental units in these cases would be the same.

Question 1: What is the sampling unit in the following cell experiment?

A technician thaws a cryo tube containing about ½ million A549 cells (Foster et al 1998) and grows the cells in a T-75 culture flask (the 75 corresponds to 75 cm^2 growing area) in a CO_2 incubator at 37 °C. After the cells reach about 70% confluency, which may represent hundreds of thousands of cells, the technician aliquots 1000 cells into twelve wells of a plate for a total of 12000 cells. To three wells the technician adds a cytokine, to another three wells he adds a cytokine inhibitor, and to another three wells he adds both the cytokine and its inhibitor, and to the last three wells he adds DMSO, which was the solvent for both the cytokine and the inhibitor. He then returns the plate to the incubator and 24 hours later extracts all of the cells and performs a multiplex quantitative PCR to determine gene expression of several target genes known to be relevant to cell proliferation.

The described experiment would be an example of a routine, but not trivial procedure the technician would do in the course of working on the project in a molecular biology laboratory.

The choices for numbers provided in the description we may consider for the number of sampling units are:

A. 1000 cells
B. 12,000 cells
C. ½ million cells
D. 3 wells
E. 12 wells
F. The target genes
G. None of the above

The correct answer is G, None of the above.

"None of the above" because the correct answer is … there was just one sampling unit! All cells trace back to that single cryo tube.

To answer the question, start from the end and work your way back. What we are looking for is independence of samples and at what level in the experiment we find independence. Our basic choice is between numbers of cells and numbers of wells. Clearly, cells are contained in the wells, so all of the cells in one well share the same medium, being treated the same way, including all the way back to the cryo tube — all of the cells came from that one tube so this lack of independence traces all the way back to the source of the cells. Thus, the answer can't be related to cell number. How many wells did the technician set up? Twelve total. So, the maximum number of sampling units would be twelve, unless the samples are not independent. And clearly the wells (samples)

are not independent because, again, all cells in the twelve wells trace back to a single cryo tube. Thus, from both perspectives, wells and cells, the answer is actually just one sampling unit! (Cumming et al 2007; Lazic 2010). Finally, the genes themselves are the targets of our study — they are the variables, not the samples. Moreover, the same logic applies — all copies of the genes are in effect descended from the few cells used to start the population.

Experimental and sampling units often, but not always the same

Question 2: What is the experimental unit in the described cell experiment?

A. 1000 cells
B. 12,000 cells
C. ½ million cells
D. 3 wells
E. 12 wells
F. The target genes
G. None of the above

The correct answer is E, 12 wells. Noted above, the technician applied treatments to 12 wells. There were two treatments, cytokine and cytokine-inhibitor (Table 5.2.1).

Table 5.2.1. Experiment description translated to a table to better visualize the design.

Well	DMSO	Cytokine	Cytokine inhibitor
1	Yes	Yes	No
2	Yes	Yes	No
3	Yes	Yes	No
4	Yes	No	Yes
5	Yes	No	Yes
6	Yes	No	Yes
7	Yes	Yes	Yes
8	Yes	Yes	Yes
9	Yes	Yes	Yes
10	Yes	No	No
11	Yes	No	No
12	Yes	No	No

Replication: Groups and individuals as sampling units

The correct identification of levels at which sampling independence occurs is crucial to successful interpretation of inferential statistics. Note the replication in Table 5.2.1: three cytokine, three cytokine-inhibitor, three with both. Sampling error rate is evaluated at the level of the sampling units. Technical replication of sampling units allows one to evaluate errors of measurement (e.g., instrument noise) (Blainey et al 2014). Replication of sampling units increases statistical power, the ability to correctly reject hypotheses. If the correct design reflects sampling units are groups and not individuals, then by counting the individuals as the independent sampling units would lead the researcher to think his design has more replication than it actually does. The consequence on the inferential statistics is that he will more likely reject a correct null hypothesis, in other words, the risk of elevated type I error occurs (Chapter 8 – Inferential Statistics). This error, confusing individual and group sampling units, is called **pseudoreplication** (Lazic 2010).

Consider a simpler experimental design scenario depicted in Figure 5.2.1: Three different water treatments (e.g., concentrations of synthetic progestins, Zeilinger et al. 2009) in bowls A, B, and C; three fish in bowl A, three fish in bowl B, and three fish in bowl C. The outcome variable might be a stress indicator, e.g., plasma cortisol (Luz et al 2008).

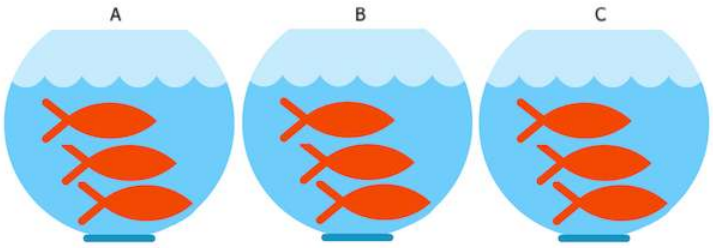


Figure 5.2.1: Three aquariums, three fish. Image modified from https://www.pngrepo.com/svg/153528/aquarium

Question 3: What were the experimental units for the fish in the bowl experiment (Fig. 5.2.1)?

A. Three bowls
B. Nine fish
C. Three water treatments

The correct answer is A, 3 bowls. The treatments were allocated to the bowls, not to individual fish. The three fish in each bowl provides technical replication for the effects of bowl A, bowl B, and bowl C, but does not provide replication for the effects of the water treatments. Adding three bowls for each water treatment, each with three fish, would be the simplest correction of the design, but may not be available to the researcher because of space or cost limitations. The design would then include nine bowls and 27 fish. If resources are not available to simply scale up the design, then the researcher could repeat the study several times, taking care to control **nuisance variables**. Alternatively, if the treatments were applied to the individual fish, then the experimental units become the individual fish and the bowls reduced to a **blocking effect** (Chapter 14.4), where differences may exist among the bowls, but they are no longer the level by which measurements are made. Note that if pseudoreplication is present in a study, this may be accounted for by specifying the error structure in a linear mixed model (e.g., random effects, blocking effects, etc., see Chapter 14 and Chapter 18).

Question 4: What were the sampling units for the fish in the bowl experiment (Fig. 5.2.1)?

A. Three bowls
B. Nine fish
C. Three water treatments

The correct answer is B, the individual fish. If instead of aqueous application of synthetic progestin, treatments were applied directly to each fish via injection, what would be the answers to Question 3 and Question 4?

Choices like these clearly involve additional compromises and assumptions about experimental design and inference about hypotheses.

Conclusion

Sampling units, experimental units, and the concept of level at which units are independent within an experiment were introduced. Lack of independence yields the problem of pseudoreplication in an experiment, which will increase the chance that we will detect differences between our treatment groups, when no such difference exists!

Questions

Figure 5.2.2: Three Miracle-Grow AeroGarden planters, each with nine seedlings of an *Arabidopsis thaliana* strain.

1. Nine seeds each of three strains of *Arabidopsis thaliana* were germinated in three Miracle-Grow AeroGarden® hydroponic planters (Fig. 5.2.2). Each planter had nine or ten vials with sphagnum peat. All seeds from a strain were planted in the same apparatus, one seed per vial. What were the experimental units?

A. planters
B. seeds
C. strains of *Arabidopsis*
D. vials in the planters

2. This experimental design is an example of pseudoreplication, but at what level?

A. planters
B. seeds
C. strains of *Arabidopsis*
D. vials in the planters

3. How would you re-do this experiment to avoid pseudoreplication? (Hint: you can't add more planters!)

5.3: Replication, bias, and nuisance

Introduction

This page describes additional elements of experimental design — replication and bias — which can impact the validity of a study.

Replication in science refers to repeating an experiment under identical conditions multiple times. To the extent the same conclusions are achieved, we can say the outcome is **reproducible** (Simmons 2014) and, therefore, likely to be **objectively true** (Earp 2011). Replication of experimental and sampling units is an essential component of good experimental design. Confusing technical replicates for biological replicates is called **pseudoreplication**.

Bias refers to any of a number of sociological and cognitive errors which may influence conclusions in science (Ionnidis 2005). In a meta-analysis, Fanelli et al (2017) reported that while the size of effect from identifiable biases were small, some disciplines were more prone to these sources of error. That researchers testing similar hypotheses reach different conclusions may result from bias. One such bias is positive outcome bias, where reports of null hypotheses are more often rejected than they should be based on sample size and other concerns of statistical power (see Chapter 11 – Power Analysis).

The page concludes with a discussion of **nuisance variables**, variables that may be causally associated with the outcome of interest, but themselves are either of no interest in the study, or may not be known in advance (Meehl 1970; see also discussion of **spurious correlations** in Chapter 16.3 – Data aggregation and correlation).

Technical vs. biological replicates

A laboratory technician retrieves a vial of an immortalized cell line from cryostorage and initiate propagation of the cells for the week's work. After a couple of days of subculturing, the technician has grown millions of cells and is ready to set up an assay. After washing the cells in PBS (phosphate buffered saline), the technician adjusts the concentration of cells to 10,000 per mL media, places the tube on ice, and proceeds to set up a 48-well microplate. To each well, the technician adds media, one or more agents, plus cells, so the total volume in each well is one mL, three replicates per treatment.

In our example there were four treatments plus a control (media only)

- DMSO only (D)
- DMSO + Cytokine (DC)
- DMSO + Small molecule Inhibitor (DS)
- DMSO + Cytokine + Small molecule Inhibitor (DCS)
- Media only (M)

The plate table might look something like Table 5.3.1.

Table 5.3.1. Schematics of a set up for a hypothetical 48-well microplate.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
A	D	D	D	◯	◯	◯	◯	◯
B	DC	DC	DC	◯	◯	◯	◯	◯
C	DS	DS	DS	◯	◯	◯	◯	◯
D	DCS	DCS	DCS	◯	◯	◯	◯	◯
E	M	M	M	◯	◯	◯	◯	◯
F	◯	◯	◯	◯	◯	◯	◯	◯

The technician has done a lot of work, but from a statistician's point of view, effectively the technician has only one sampling unit to show for his efforts; there is one biological sampling unit because as far as we can tell, the cells were all related. **Biological replication** refers to repeat measures (e.g., same conditions) on biologically distinct entities (Blainey et al 2014). **Technical replication**, in contrast, refers to repeat measures on the same entity, e.g., the same individual (Blainey et al 2014).

However, there were more than one technical samples in the experimental design: twelve in all, because there were twelve wells. Many of you will recognize this as the distinction between technical and biological replicates. In thinking about experimental

design, replicates pertain to the number of sampling units that are treated the same way. In our example cells may be "treated the same way" in at least two distinct ways. Biologically, the cells may be the same: derived from the same clonal cell, all at the same passage (generation) number, each cell genetically, morphologically, physiologically, etc. identical to the next cell. Technical replication on the other hand pertains to the levels of the treatments.

Within each of these treatments we had 3 wells, and that's at the level of the technical replication.

A final note on this experiment. We as molecular biologists call each of these treatment groups, but statisticians thinking of this design would recognize only one treatment (called in statistics a factor), with four levels of the treatment. We will discuss this beginning in chapter 12, when we introduce analysis of variance.

Types of bias

When we make decisions, we like to think we are rational; that we make decisions based on an evaluation of evidence. And yet, an increasing body of literature suggests that our decisions often made in a manner that falls short of rational processing. We outlined some sources of bias in Chapter 2.6. Deviance from rational decision making is due to any number of **cognitive biases** we may have. Researchers and medical doctors make many decisions and, unfortunately, are just as susceptible to cognitive biases as the rest of us. One kind of bias that is bedeviling to research is **confirmatory bias**. Confirmatory bias refers to a researcher in effect seeking evidence in a way that confirms a prior conclusion (Kaptchuck 2003). Confirmatory bias would be exhibited if we report positive effects of aspirin to relieve migraines in subjects, but exclude cases where the subjects report no improvement. Medical doctors may be risk-averse and therefore tend to over-prescribe, or they may be risk-takers (e.g., rise and fall of high-dose chemotherapy, cf. discussion in Howard et al. 2011).

The concept of bias was first introduced in Chapter 3 and will be returned to in subsequent chapters. In general, bias refers to the objectivity of measurement and inferences about such measurements. Bias implies that a series of measurements consistently fail to return the true value of the variable. Bias is systematic error and may be associated with a poorly calibrated instrument or even the use of improper sets of rules for measurement given the nature of the characteristic. Bias can be captured by the concept of **accuracy**.

Bias is challenging to eliminate; the best way is to design the experiment so that the observer, the researcher, is unaware of the specific questions to be tested. The research is conducted "blind." Clinical research provides the easiest examples. A trial can be blind in a couple of ways:

- The subjects do not know which treatment they receive, but the researcher does know
- The researcher does not know which treatment subjects receive, but the subjects know
- Neither the researcher nor the subjects know which treatment they receive.
- The other possibility is that both the researcher and the subjects know treatments received — this would be a poor design.

Bias can enter a research project at multiple levels. A partial list of sources of bias in research includes (modified after Pannucci and Wikins 2010):

Bias that occurs before the experiment:

- Inadequate planning of the experiment, flawed design
- Bias in selection of subjects from the reference population
- Surveillance bias, where one group is studied more closely than another group.

Performance bias refers to conditions of an experiment that introduce unintended differences between groups. For example, subjects enrolled in a weight-loss study randomly assigned to the control group may react poorly when they realize they are not receiving the experimental intervention. This leads to potential for a systematic bias — participants in the control group may behave differently, counteracting the point of randomization (McCambridge et al 2014).

Another well-known performance bias is associated with eliciting maximal performance from animals, such as running stamina or maximum sprint running speed (Losos et al 2002). For example, spring running speed of lizards may be measured by placing lizards onto a high-speed treadmill, then increasing belt speed to match the individual's burst performance (e.g., Dohm et al 1998). Clearly, not all individuals will perform to maximum physiological capabilities under these conditions.

Bias after the experiment:

- Citation bias
- Confounding

Another way to view sources of bias is that, at least from the perspective of the researcher, the bias is likely to be **unconscious bias**. After all, if we knew about the sources of bias we would work our experiments in such a way to minimize known error-producing sources (Holman et al 2015). Selecting subjects at random and keeping record keeping blind with respect to treatments are among the best tools we have to avoid this kind of bias.

Nuisance variables

A key quality of experimental design is the opportunity afforded the researcher to apply control. In a classic view of the experiment, one allows only one aspect to vary, the treatment applied, all else is controlled. Thus, any difference in outcome may be attributed to the treatment. However, it is likely impossible to control for all possible variables that may effect our outcome; **confounding** clear interpretation of results (see Chapter 16.2 and 16.3). Consider, for example, the effects of age on running performance (Fig. 5.3.2).

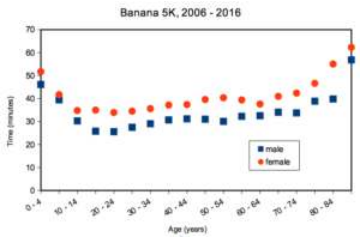


Figure 5.3.2: Mean 5K running times (minutes) by age and gender (2006 – 2016, Jamba Juice Banana 5K race, Honolulu, HI).

If we are interested in gender differences, we must also deal with the change in performance with age: older runners regardless of gender tend to run slower than younger runners. Thus, in the comparison of genders, age is a *nuisance variable*. In the literature, you may see "lurking variable" or other such names used. The concept is that these variables may be causal, but are either of no interest in the study, or may not be known in advance. Therefore, age must be accounted for before we can address gender differences, if any. Age, in this case, is considered a nuisance variable because we are not primarily interested in age effects, but age clearly covaries (is associated) with running performance. How best to handle nuisance variables? One approach is to match by age (**blocking effect**, see Chapter 14.4); another approach would be to randomly select with respect to age when compiling female and male times.

Here is another example of confounding. Results from ecological experiments intended to test for presence of competition for resources by plant species by removing plants may alter herbivory levels; thus, survivorship may not be the result of reduced plant-plant competition, but the result of changes in herbivore behavior (Reader 1992).

Questions

1. I cite a number of articles in Mike's Biostatistics Book. If you're paying attention, you'll have noticed that all of my citations, at least for articles, link to a source that you can either download or read online. What kind of error am I committing by following this approach?
2. A doctor has been seeing patients with upper respiratory tract infections (URTI) all week. Although most URTI are caused by viral infections (Thomas and Bomar 2018), the doctor has prescribed each patient a ten-day dose of antibiotics. As you should know, antibiotics have no effect on the course of viral infections; antibiotics work against bacterial infections. What kind of error has the doctor committed by following this routine prescription of antibiotics?
3. Figure 5.3.2 is a plot of mean running times for men and women across different ages. What statistic is missing from the graph so that we can't conclude from the graph anything more than there is a trend that men are faster than women?
4. In large part because of the tendency for over use of antibiotics, doctors are less likely today to prescribe antibiotics for patients with upper respiratory tract infections than in the past (Zoorob et al 2012). URTI are mostly caused by rhinovirus infections and, therefore, the course of infection in a patient should be unaffected by addition of antibiotic. However, although rare compared to URTI, life-threatening illnesses like acute epiglottis caused by bacteria infection can sometimes develop or accompany URTI. Thus, not prescribing antibiotics for a diagnosis of URTI may risk a worsening condition for the patient. This sets the doctor up for a potential cognitive bias – to prescribe or not to prescribe antibiotics prophylactically? Discuss this decision in the context of potential sources of cognitive biases and possible outcomes of the decision.

5.4: Clinical trials

Introduction

Two areas of inquiry have contributed enormously to our understanding of experimental design: agriculture research and human subject **clinical trials**. This page outlines types of clinical trial designs and briefly introduces the subject of research ethics with respect to human subject research.

An outline of clinical trials

Much has been written about biomedical research study design; a couple of accessible articles that can supplement the material presented here include Benson and Hart 2000, Concato et al 2000, Gabriella 2012. Even if you never work in clinical research, understanding how clinical trials are designed and under what circumstances limitations of particular designs arise is helpful to all of us who do experimental work. In Mike's Workbook for Biostatistics we will present descriptions of research activities and step through situations where you will attempt to align the research by analogy to clinical trials. A second approach to learning about experimental design is to read about someone's study, and reworking it in terms of a clinical trial design perspective. We will use a discussion of clinical trials, used in biomedical research to investigate effectiveness of treatments of disease, as our starting point for learning how statistics informs experimentation.

Types of clinical trials are distinguished by their design and include:

Experimental studies are just that, research designs that apply techniques of experimental science — controls, randomization, attempts to account for sources of bias. They are intended to make direct comparisons among subjects assigned by the researcher to treatment groups. By definition experiments are longitudinal studies. **Longitudinal studies** are experimental or observational studies in which multiple observations are recorded for each individual and individuals are tracked over time. Many excellent resources about experimental design are available, from R.A. Fisher's 1935 book, *The design of experiments*, to Scheiner and Gurevitch (Editors) 2001, *Design and Analysis of Ecological Experiments* (2nd edition), to the many books on randomized control trials, e.g., the 5th edition of *Designing clinical research* (2022), edited by Browner et al.

Observational (or epidemiological) **studies** — no direct intervention is administered and so observational studies tend to be **retrospective**; we identify individuals with and without the condition and attempt to identify associations between the condition and any number of potential causal factors.

Cross-sectional studies are examples of **descriptive study** designs. They take observations at one point in time on a variety of individuals. It can be used to associate factors with the condition in question, and can be used to estimate the prevalence of a condition in the population. Cross-sectional studies are referred to as `method = cross.sectional` in the package epiR. Omair (2015) provides an accessible summary of case and cross-sectional study designs.

Cohort studies involve a group of subjects (e.g., patients) who receive the same treatment at the same time. Cohort studies are referred to as `method = cohort.count` in the package epiR. A cohort consists of subjects who are linked in some way. It could be a trivial link, like the cohorting done at university (all incoming Freshmen students who enroll for a class offered at 9:30 AM), or it could be based on shared experience due to an exposure event (e.g., all passengers on a jet traveling with an index case or "patient zero").

- Prospective cohort studies enroll people as cohorts at the beginning of a study and follow them over time.
- Retrospective cohort studies may utilize archived records.

Cohort and other variations of observational studies (e.g., **case control**) can establish associations between risk factors and conditions or specific adverse events, but cannot by themselves establish cause and effect (Benson and Hartz 2000).

Case control studies are similar to cohort studies, except they are retrospective. Case refers to subjects with one or more characteristics of interest. Used to infer the exposure risk factor by evaluating historical records. Case control studies are referred to as `method = case.control` in the package epiR. Omair (2016) provides an accessible summary of cohort and case control study designs. Designed to identify associations between exposure and particular outcomes, case control studies are retrospective and observational studies: retrospective because the outcomes are already known, and observational because the event was caused by nature, not experiment. In principle, researchers identify a number of cases with a particular outcome (e.g., lung cancer), then attempt to match cases to individuals who do not have the outcome (controls). Work is done to look back to see if the exposure (e.g., smoking) is more frequent in the case group than the control group. Case control studies have several advantageous compared to other approaches: they are rapid to conduct compared to longitudinal studies (the event has already happened), and efficient

because small sample sizes may be enough to reach conclusions. Among their limitations, however, is the problem of how to match cases with controls. Obvious matching accomplished by grouping by age categories, body mass index, gender, socio-economic status, and so on. Three papers authored by Wacholder published in American Journal of Epidemiology 1992 describe in detail, from theory to practice, case control selection (Wacholder 1992abc).

Randomized Control (interventional or experimental) **Trials** (RCT): compares an experimental treatment group with a control) placebo group. The groups are assigned to groups randomly. Another variant of a RCT is a randomized clinical trial, with the only difference that the clinical trial compares different treatments and may not include a control group.

Double-blind: The "gold standard" RCT. Both the patient and those interacting with the patient do not know what treatment the patient has received.

Placebo: a treatment in name only. The placebo is a designed but medically ineffective agent given to a study subject. There is considerable debate about what a placebo should contain and as to the ethics and general merits of its use in clinical trials (Temple and Ellenberg 2000).

- **Active control** or **comparator** studies are now common in place of placebo studies where treatment is clearly better than not giving the subject something of benefit (i.e., the placebo which is designed to not benefit the subject). Active control is not automatically a better choice than placebo control, because such studies may be less effective in evaluating cause and effect (Temple and Ellenberg 2000).

The controlled clinical trial has a long history: **Daniel's vegetarian diet** (Daniel's training in Babylon, Book of Daniel, Old Testament, discussed in Bhatt 2010, h/t Treece 1990) — after ten days those on Daniel's diet looked healthier than the others who ate the King's prescribed meal of meat and wine; James Lind's **scurvy trials** on board *HMS Salisbury*, a British ship in 1747 (references in Bhatt 2010). Randomized control trials (RCT) were introduced by Hill and others in a 1946 study of streptomycin efficacy against tuberculosis. The effectiveness of RCT is now established and integral to regulation of drug development; see Figure 5.4.1. Use of clinical trials, unfortunately, has had a longer history than recognition of human rights. World War II Nazi medical research atrocities are well known (Berger 1990), so too the longitudinal study of Tuskegee syphilis study on African-Americans (Brandt 1978). But there are too many other examples: in 1884 Hawaii, the inoculation of Keanu by Dr. Arning with **leprosy** in exchange for commuting Keanu's death sentence to life imprisonment (Keanu developed leprosy and died in 1890, Binford 1936); many studies of American Indians/Alaskan natives (Hodge 2012). Rules of conduct were established at Nuremberg, and subsequently extended and codified by the Belmont Accords: core principles of respect for persons, beneficence, and justice. Ethical standards of who participates are institutionalized by IRB boards. True informed consent remains challenging (Rothwell et al 2021 and references therein).

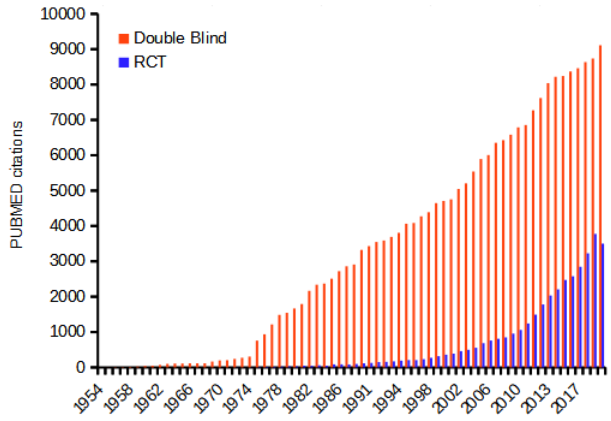


Figure 5.4.1: The rising number of PUBMED citations for double blind and RCT studies, by year from 1954 to 2020.

Ethics of clinical and experimental research

Informed consent of subjects before proceeding in a clinical trial is a required, essential component of the design of a clinical trial. Guidelines for research conducted on human subjects originated from The **Nuremberg Code** (Shuster 1997). The Code was formulated 67 years ago, in August 1947, in Nuremberg, Germany, by American judges sitting in judgment of Nazi doctors accused of conducting torturous experiments on humans in the concentration camps (Shuster 1997). It served as a blueprint for today's principles that ensure the rights of subjects in medical research. Achieving informed consent is not always straightforward (Nijhawan et al 2013), and we continue to see research that challenges ethical standards (e.g., discussion in Suba 2014).

Additionally, and perhaps less appreciated, the Nuremburg Code is justification for invasive animal research: animal research must precede human subject testing (Shuster 1997).

While informed consent is required, clearly, it is not enough. Emmanuel et al (2000) provide a framework for evaluating the ethics of a research program involving human subjects:

1. Value
2. Scientific validity
3. Fair subject selection
4. Favorable risk-benefit ratio
5. Independent review
6. Informed consent
7. Respect for enrolled subjects

An additional concept, **clinical equipoise** (Freedman 1987), is relevant. Freedman noted that the researcher must have "genuine uncertainty" with respect to the merits of each treatment, or an "honest null hypothesis." If a consensus exists that one treatment is better than another, including placebo, then there is no null hypothesis and the research would be invalid (Emmanuel et al 2000). Take, for example, the suggestion that clinicians should withhold angiotensin-converting inhibitors (ACE2) from their hypertensive Covid-19 patients (Fang et al 2020; discussed in Tignaneli et al 2020). The hypothesis comes from the observation that SARS-COV2, like SARS-Cov, binds with ACE2 receptor in order to invade the cell. Blocking ACE2 inhibitors then would reduce activation of pulmonary renin angiotensin system and subsequent lung injury. Tignaneli et al (2020) called this a case of clinical equipoise — they argued no evidence supports "routine discontinuation" of ACE inhibitors.

Guidelines mandate detail about experimental design

Professional journals expect authors to provide detailed descriptions of all methodology, including aspects of experimental design. Fundamental to the aim of science, to increase our knowledge An essential component of science To improve this kind of communication many journals and professional societies have promoted standards about what must be included in these descriptions. For example, efforts of the **CONSORT**, which stands for **CON**solidated **S**tandards **O**f **R**eporting **T**rials (www.consort-statement.org), to improve the reporting of clinical trials by authors and provide guidelines for reviewers and editors are endorsed by more than 400 journals. The CONSORT checklist addresses

- Trial design
- Participants
- Interventions
- Outcomes
- Sample size
- Randomization
- Implementation
- Blinding
- Statistical methods

These nine elements were judged essential for authors to report how their study implemented or did not implement. The purpose of these items is to in order to improve **reproducibility of published research**. For **animal research**, a similar list is available from **ARRIVE** (Kilkenny et al 2010). ARRIVE also addresses additional criteria and directs how these should be reported throughout the paper, not just in the methods section. Like CONSORT, hundred of journals have endorsed the ARRIVE 20-item checklist.

Clinical researchers must implement protocols to insure data management guidelines are followed. **Clinical data management** is a large topic in and of itself, so we won't discuss this area further. However, good data management practice across disciplines share a number of features. For example, all data records should include metadata, where **metadata** refers to "information about data," and would include enough information about the experiment, including

- dates and times of observations
- personnel
- facilities
- protocols used
- number of subjects
- list of variables with definitions

- full name of variable plus any acronym
- measurement units
- instrumentation
- notes about data quality

- conditions

and more (this is hardly an exhaustive list). Metadata therefore explains how data were obtained. Lists of variables are also called **data dictionaries**. If data are stored in spreadsheets, for example, then good practice includes including a worksheet with the metadata for the data set.

Questions

1. Be able to define the following terms:
   - case control
   - case study
   - cohort study
   - cross sectional study
   - observational study
   - experimental study
   - single arm trial
   - single blind vs double-blinding in research design
2. Early in the Covid-19 pandemic, hydroxycholoquine was suggested for treating Covid-19 patients, and some called for prophylactic use of the malarial drug. Discuss the treatment hypothesis in the context of clinical equipoise.
3. Distinguish between case control prospective and case control retrospective studies, and the kinds of inferences that can be made from each.

5.5: Importance of randomization

Introduction

If the goal of the research is to make general, evidenced-based statements about causes of disease or other conditions of concern to the researcher, then how the subjects are selected for study directly impacts our ability to make **generalizable conclusions**. The most important concept to learn about inference in statistical science is that your sample of subjects, upon which all measurements and treatments are conducted, ideally should be a random selection of individuals from a well-defined reference population.

The primary benefit of **random sampling** is that it strengthens our confidence in the links between cause and effect. Often after an **intervention trial** is complete, differences among the **treatment groups** will be observed. Groups of subjects who participated in sixteen weeks of "vigorous" aerobic exercise training show reduced systolic blood pressure compared to those subjects who engaged in light exercise for the same period of time (Cox et al 1996). But how do we know that exercise training *caused* the difference in blood pressure between the two treatment groups? Couldn't the differences be explained by chance differences in the subjects? Age, **body mass index** (BMI), overall health, family history, etc.?

How can we account for these additional differences among the subjects? If you are thinking like an experimental biologist, then the word "control" is likely coming to the foreground. Why not design a study in which all 60 subjects are the same age, the same BMI, the same general health, the same family history…? Hmm. That does not work. Even if you decide to control age, BMI, and general health categories, you can imagine the increased effort and cost to the project in trying to recruit subjects based on such narrow criteria. So, control per se is not the general answer.

If done properly, random sampling makes these alternative explanations less likely. Random sampling implies that other factors that may causally contribute to differences in the measured outcome, but themselves are not measured or included as a focus of the research study, should be the same, on average, among our different treatment groups. The practical benefits of proper random sampling is that recruiting subjects gets easier — fewer subjects will be needed because you are not trying to control dozens of factors that may (or may not!) contribute to differences in your outcome variable. The downside to random sampling is that the variability of the outcomes within your treatment groups will tends to increase. As we will see when we get to statistical inference, large variability within groups will make it less likely that any statistical difference between the treatment groups will be observed.

Demonstrate the benefits of random sampling as a method to control for extraneous factors.

The study reported by Cox et al. included 60 obese men between the ages of 20 and 50. A reasonable experimental design decision would suggest that the 60 subjects be split into the two treatment groups such that both groups had 30 subjects for a balanced design. Subjects who met all of the research criteria and who had signed the informed consent agreement are to be placed into the treatment groups and there are many ways that group assignment could be accomplished. One possibility, the researchers could assign the first 30 people that came into the lab to the Vigorous exercise group and the remaining 30 then would be assigned to the Light exercise group. Intuitively I think we would all agree that this is a suspect way to design an experiment, but more importantly, why shouldn't you use this convenient method?

Just for argument's sake, imagine that their subjects came in one at a time, and, coincidentally, they did so by age. The first person was age 21, the second was 22, and so on up to the 30th person, who was 50. Then, the next group came in, again, coincidentally in order of ascending age. If you calculate the simple average age for each group you will find that they are identical (35.5 years). On the surface, this looks like we have controlled for age: both treatment groups have subjects that are the same age. A second option is to sort the subjects into the two treatment groups so that one 21-year-old is in Group A and the other 21-year-old is in Group B, and so on. Again, the average age of Group A subjects and of Group B subjects would be the same and therefore controlled with respect to any covariation between age and change in blood pressure. However, there are other variables that may covary with blood pressure, and by controlling one, we would need to control the others. Randomization provides a better way.

I will demonstrate how randomization tends to distribute the values in such a way that the groups will not differ appreciably for the **nuisance variables** like age and BMI differences and, by extension, any other covariable. The R work is attached following the Reading list. The take-home message: After randomly selecting subjects for assignment to the treatment groups, the apparent differences between Group A and Group B for *both* age and BMI are substantially diminished. No attempt to match by age and by BMI is necessary. The numbers are shown in the table and then in two graphics (Fig. 5.5.1, Fig. 5.5.2) derived from the table.

Table 5.5.1. *Mean age and BMI for subjects in two treatment groups A and B where subjects were assigned randomly or by convenience to treatment groups.*

	Group	Random assignment of subjects to treatment groups	Convenience assignment of subjects to treatment groups
Mean age	A	35.2	28
	B	35.8	43
Mean BMI	A	32.49	28.99
	B	32.87	37.37

Just for emphasis, the means from Table 5.5.1 are presented in the next two figures (Fig. 5.5.1 and Fig. 5.5.2).

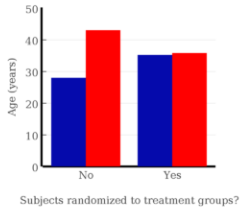


Figure 5.5.1: Age of subjects by groups (A = blue, B = red) with and without randomized assignment of subjects to treatment groups.

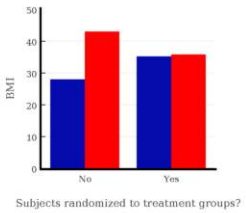


Figure 5.5.2: BMI of subjects by groups (A = blue, B = red) with and without randomized assignment of subjects to treatment groups.

Note that the apparent difference between A and B for BMI disappears once proper randomization of subjects was accomplished. In conclusion, a random sample is an approach to experimental design that helps to reduce the influence other factors may have on the outcome variable (e.g., change in blood pressure after 16 weeks of exercise). In principle, randomization should protect a project because, on average, these influences will be represented randomly for the two groups of individuals. This reasoning extends to unmeasured and unknown causal factors as well.

This discussion was illustrated by random assignment of subjects to treatment groups. The same logic applies to how to select subjects from a population. If the sampling is large enough, then a random sample of subjects will tend to be representative of the variability of the outcome variable for the population and representative also of the additional and unmeasured cofactors that may contribute to the variability of the outcome variable.

What about observational studies? How does randomization work?

However, if you do cannot obtain a random sample, then conclusions reached may be sample-specific, **biased**… perhaps the group of individuals that likes to exercise on treadmills just happens to have a higher cardiac output because they are larger than the individuals that like to exercise on bicycles. This nonrandom sample will bias your results and can lead to incorrect interpretation of results. Random sampling is CRUCIAL in epidemiology, opinion survey work, and most aspects of health, drug studies, medical work with human subjects. It's difficult and very costly to do… so most surveys you hear about, especially polls reported from Internet sites, are NOT conducted using random sampling (included in the catch-all term "**probability sampling**")!! As an aside, most opinion survey work involves complex sample designs involving some form of **geographic clustering** (e.g., all phone numbers in a city, random sample among neighborhoods).

Random sampling is the ideal if generalizations are to be made about data, but strictly random sampling is not appropriate for all kinds of studies. Consider the question of whether or not EMF exposure is a risk factor for developing cancer (Pool 1990). These kinds of studies are observational: at least in principle, we wouldn't expect that housing and therefore exposure to EMF is manipulated (cf. discussion Walker 2009). Thus, epidemiologists will look for patterns: if EMF exposure is linked to cancer, then more cases of cancer should occur near EMF sources compared to areas distant from EMF sources. Thus, the hypothesis is that an association between EMF exposure and cancer occurs non-randomly, whereas cancers occurring in people not exposed to EMF are random. Unfortunately, clusters can occur even if the process that generates the data is random.

Compare Graph A and Graph B (Fig. 5.5.3). One of the graphs resulted from a **random process** and the other was generated by a **non-random process**. Note that the claim can be rephrased about the probability that each grid has a point, e.g., it's like Heads/Tails of 16 tosses of a coin. Which graph shows a randomly generated data set? We can see clusters of points in Graph B; Graph A lacks obvious clusters of points — there is a point in each of the 16 cells of the grid. Although both patterns could be random, the correct answer in this case is Graph B.

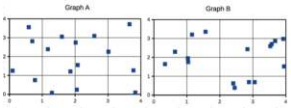


Figure 5.5.3: An example of clustering resulting from a random sampling process (Graph B). In contrast, Graph A was generated so that a point was located within each grid.

The graphic below shows the transmission grid in the continental United States (Fig. 5.5.4). How would one design a random sampling scheme overlaid against the obviously heterogeneous distribution of the grid itself? If a random sample was drawn, chances are good that no population would be near a grid in many of the western states, but in contrast, the likelihood would increase in the eastern portion of the United States where the population and therefore transmission grid are more densely placed.

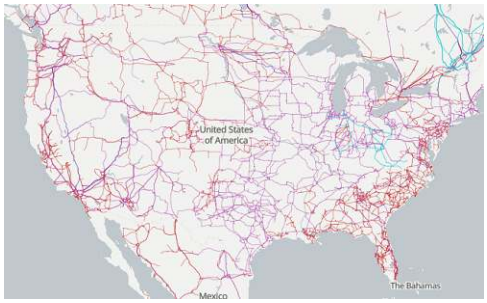


Figure 5.5.4: Map of electrical transmission grid for continental United States of America. Image source https://openinframap.org/#3/24.61/-101.16.

For example, you want to test whether or not EMF affects human health, and your particular interest is in whether or not there exists a relationship between brain cancer and living close to high voltage towers or transfer stations. How does one design a study, keeping in mind the importance of randomization for our ability to generalize and assign causation? This is a part of epidemiology which strives to detect whether clusters of disease are related to some environmental source. It is an extremely difficult challenge. For the record, no clear link to EMF and cancer has been found, but reports do appear from time to time (e.g., report on a cluster of breast cancer in men working in office adjacent to high EMF, Milham 2004).

Questions

1. I claimed that Graph B in Figure 5.5.3 was generated by a random process while Graph B was not. The results are: Graph A, each cell in the grid has a point; In graph B, ten cells have at least one point, six cells are empty. Which probability _____ distribution applies?

A. beta
B. binomial
C. normal
D. poisson

2. Confirm the claim by calculating the probability of Graph A result vs Graph B result.

R code!

Recall that statements preceded by the hash # are comments and are not read by R (i.e., no need for you to type them).

First, create some variables. Vectors aa and bb contain my two age sequences.

```
aa = seq(21,50)
bb = seq(21,50)
```

Second, append vector bb to the end of vector aa

```
age = append(aa,bb)
age           #submit the vector (variable) name to print the records for verification - looks good!
 [1] 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 21 22 23 24
[48] 38 39 40 41 42 43 44 45 46 47 48 49 50
```

Third, get the average age for the first group (the `aa` sequence) and for the second group (the `bb` sequence). Lots of ways to do this: I made two subsets from the combined age variable; I could have just as easily taken the mean of `aa` and the mean of `bb` (same thing!).

```
A = age[1:30]
mean(A)
[1] 35.5
B = age[31:60]
mean(B)
[1] 35.5
```

Fourth, start building a data frame, then sort it by age. We will be adding additional variables to this data frame.

```
ex.random = data.frame(age)
AO.all = sort(ex.random$age)
AO.age           #submit the vector (variable) name to print the records for verification - looks good!
 [1] 21 21 22 22 23 23 24 24 25 25 26 26 27 27 28 28 29 29 30 30 31 31 32 32 33 33 34 34 35 35 36 36 37 37
[48] 44 45 45 46 46 47 47 48 48 49 49 50 50
```

Fifth, divide the variable again into two subsets of 30 and get the averages.

```
AO = AO.age[1:30]
AO
 [1] 21 21 22 22 23 23 24 24 25 25 26 26 27 27 28 28 29 29 30 30 31 31 32 32 33 33 34 34 35 35
mean(AO)
[1] 28
BO = AO.age[31:60]
BO
 [1] 36 36 37 37 38 38 39 39 40 40 41 41 42 42 43 43 44 44 45 45 46 46 47 47 48 48 49 49 50 50
mean(BO)
[1] 43
```

Sixth, create an index variable, random order without replacement.

```
rand.index = sample(1:60,60,replace=F)
```

Add the new variable to our existing data frame, then print it to check that all is well.

```
ex.random$rand = rand.index
ex.random
   age rand
1   21   43
2   22   15
3   23   17
4   24   35
5   25   19
6   26   18
7   27   22
8   28   31
9   29   12
10  30   44
11  31   24
12  32    5
13  33    2
14  34   50
15  35   23
16  36   20
17  37   41
18  38   56
19  39   36
20  40    8
21  41   45
22  42   38
23  43   42
24  44   46
25  45   16
26  46   21
27  47   28
28  48   10
29  49   32
30  50   54
31  21   57
32  22   51
33  23   27
34  24   40
35  25   14
36  26   48
37  27   26
38  28   58
39  29    9
40  30   11
41  31    4
42  32   52
43  33   37
44  34   53
45  35    6
46  36   34
47  37   39
48  38    7
49  39    1
50  40   47
51  41   33
52  42   60
53  43   49
54  44   30
55  45   29
56  46   55
57  47   13
58  48    3
59  49   25
60  50   59
```

Seventh, select for our first treatment group the first 30 subjects from the randomized index. There are again other ways to do this, but sorting on the index variable means that the subject order will be change too.

```
AR.age = ex.random[order(ex.random$rand),]     #created a new data frame to distinguish it from the presor
```

Print the new data frame to confirm that the sorting worked. It did. We can see that the rows have been sorted by ascending order based on the index variable.

```
AR.age
   age rand
49  39    1
13  33    2
58  48    3
41  31    4
12  32    5
45  35    6
48  38    7
20  40    8
39  29    9
28  48   10
40  30   11
9   29   12
57  47   13
35  25   14
2   22   15
25  45   16
3   23   17
6   26   18
5   25   19
16  36   20
26  46   21
7   27   22
15  35   23
11  31   24
59  49   25
37  27   26
33  23   27
27  47   28
55  45   29
54  44   30
8   28   31
29  49   32
51  41   33
46  36   34
4   24   35
19  39   36
43  33   37
22  42   38
47  37   39
34  24   40
17  37   41
23  43   42
1   21   43
10  30   44
21  41   45
24  44   46
50  40   47
36  26   48
53  43   49
14  34   50
32  22   51
42  32   52
44  34   53
30  50   54
56  46   55
18  38   56
31  21   57
38  28   58
60  50   59
52  42   60
```

Eighth, create our new treatment groups, again of n = 30 each, then get the mean ages for each group.

```
AR = AR.age$age[1:30]
mean(AR)
[1] 35.16667
AR2 = AR.all$all[31:60]
mean(AR2)
[1] 35.83333
```

Get the minimum and maximum values for the groups

```
min(AR)
[1] 22
min(AR2)
[1] 21
max(AR)
[1] 49
max(AR2)
[1] 50
```

Ninth, create a BMI variable drawn from a normal distribution with coefficient of variation equal to 20%. The first group we will call cc.

```
cc = rnorm(n=30,m=27.5, sd=5.5)          #mean was 27.5 for this group with standard deviation of 5.5
```

The second group will be called dd.

```
dd = rnorm(n=30,m=37.5, sd=7.5)          #mean was 37.5 for this group with standard deviation of 7.5
```

Create a new variable called BMI by joining cc and dd.

```
BMI=append(cc,dd)
BMI            #print out BMI to confirm. Looks good!
 [1] 27.87528 27.83250 31.88703 34.99041 24.06751 23.50952 22.57779 31.48394 31.04321 25.60258 25.41081 22
[16] 20.56529 27.25238 21.85205 32.11690 32.37168 23.11314 33.29110 34.99106 38.22016 18.72105 26.22030 25
[31] 47.57872 27.58428 40.17211 38.22195 26.91893 37.02784 53.72671 34.94727 30.35245 38.32571 40.52111 36
[46] 40.30846 36.47643 50.86804 43.63741 37.84994 42.82665 41.71008 28.44976 24.57906 42.37762 38.38512 35
```

Add the BMI variable to our data frame.

```
ex.random$BMI = BMI
ex.random              #Print out the revised data frame. Looks good. We now have three variables: age, th
   age rand      BMI
1   21   43 27.87528
2   22   15 27.83250
3   23   17 31.88703
4   24   35 34.99041
5   25   19 24.06751
6   26   18 23.50952
7   27   22 22.57779
8   28   31 31.48394
9   29   12 31.04321
10  30   44 25.60258
11  31   24 25.41081
12  32    5 22.34619
13  33    2 34.62213
14  34   50 36.41348
15  35   23 41.17740
16  36   20 20.56529
17  37   41 27.25238
18  38   56 21.85205
19  39   36 32.11690
20  40    8 32.37168
21  41   45 23.11314
22  42   38 33.29110
23  43   42 34.99106
24  44   46 38.22016
25  45   16 18.72105
26  46   21 26.22030
27  47   28 25.13412
28  48   10 27.50475
29  49   32 34.79361
30  50   54 32.81267
31  21   57 47.57872
32  22   51 27.58428
33  23   27 40.17211
34  24   40 38.22195
35  25   14 26.91893
36  26   48 37.02784
37  27   26 53.72671
38  28   58 34.94727
39  29    9 30.35245
40  30   11 38.32571
41  31    4 40.52111
42  32   52 36.15627
43  33   37 30.36592
44  34   53 36.20397
45  35    6 47.63142
46  36   34 40.30846
47  37   39 36.47643
48  38    7 50.86804
49  39    1 43.63741
50  40   47 37.84994
51  41   33 42.82665
52  42   60 41.71008
53  43   49 28.44976
54  44   30 24.57906
55  45   29 42.37762
56  46   55 38.38512
57  47   13 35.22879
58  48    3 31.34063
59  49   25 34.02996
60  50   59 27.28038
```

Tenth, repeat our protocol from before: Set up two groups each with 30 subjects, calculate the means for the variables and then sort by the random index and get the new group means.

```
AO = ex.random$BMI[1:30]
mean(AO)
[1] 28.99333
BO = ex.random$BMI[31:60]
mean(BO)
[1] 37.36943
```

All we did was confirm that the unsorted groups had mean BMI of around 27.5 and 37.5 respectively. Now, proceed to sort by the random index variable. Go ahead and create a new data frame.

```
AR.age = ex.random[order(ex.random$rand),]
AR.age                 #Print out the new data frame to confirm. Looks good.
   age rand      BMI
49  39    1 43.63741
13  33    2 34.62213
58  48    3 31.34063
41  31    4 40.52111
12  32    5 22.34619
45  35    6 47.63142
48  38    7 50.86804
20  40    8 32.37168
39  29    9 30.35245
28  48   10 27.50475
40  30   11 38.32571
9   29   12 31.04321
57  47   13 35.22879
35  25   14 26.91893
2   22   15 27.83250
25  45   16 18.72105
3   23   17 31.88703
6   26   18 23.50952
5   25   19 24.06751
16  36   20 20.56529
26  46   21 26.22030
7   27   22 22.57779
15  35   23 41.17740
11  31   24 25.41081
59  49   25 34.02996
37  27   26 53.72671
33  23   27 40.17211
27  47   28 25.13412
55  45   29 42.37762
54  44   30 24.57906
8   28   31 31.48394
29  49   32 34.79361
51  41   33 42.82665
46  36   34 40.30846
4   24   35 34.99041
19  39   36 32.11690
43  33   37 30.36592
22  42   38 33.29110
47  37   39 36.47643
34  24   40 38.22195
17  37   41 27.25238
23  43   42 34.99106
1   21   43 27.87528
10  30   44 25.60258
21  41   45 23.11314
24  44   46 38.22016
50  40   47 37.84994
36  26   48 37.02784
53  43   49 28.44976
14  34   50 36.41348
32  22   51 27.58428
42  32   52 36.15627
44  34   53 36.20397
30  50   54 32.81267
56  46   55 38.38512
18  38   56 21.85205
31  21   57 47.57872
38  28   58 34.94727
60  50   59 27.28038
52  42   60 41.71008
```

Get the means of the new groups.

```
AR = AR.age$BMI[1:30]
mean(AR)
[1] 32.49004
min(AR)
[1] 18.72105
max(AR)
[1] 53.72671
AR2 = AR.all$BMI[31:60]
mean(AR2)
[1] 33.87273
min(AR2)
[1] 21.85205
max(AR2)
[1] 47.57872
```

That's all of the work!

5.6: Sampling from populations

Introduction

Researchers generally can't study an entire **population**. More generally, striving to study each member of a population is not necessary to arrive at answers about the population. For example, consider this question: does taking a multivitamin daily improve health? What are our options? Do we really need to follow every single individual in the United States of America, monitoring their health and noting whether or not the person takes vitamins daily in order to test (**inference**) this hypothesis? Or, can we get at the same answer by careful experimental design (see Dawsey et al 2014)? Supplement use is widespread in the United States, but both health and vitamin use differ by demographics. Young people tend to be healthier than older people and older people tend to take supplements more than younger people.

A subset of the population is measured for some trait or characteristic. From the sample, we hope to refer back to the population. We want to move from anecdote (case histories) to possible generalizations of use to the reference population (all patients with these symptoms). How we sample from the reference population limits our ability to generalize. We need a representative sample: simple to define, hard to achieve.

Statistics becomes necessary if we want to infer something about the entire populations. (Which is usually the point of doing a study!!) Typically, tens to thousands of individuals are measured. But in addition, HOW we obtain the sample of individuals from the reference population is CRITICAL.

Kinds of sampling include (adapted from Box 1, Tyrer and Heyman 2016):

- Probability sampling
  - random
  - stratified
  - clustered
  - systematic
- Nonprobability sampling
  - convenience, haphazard
  - judgement
  - quota
  - snowball

How can samples be obtained?

Sampling from a population may be convenient. For one famous example, consider the Bumpus data set. (We introduced this data set in Question 5, Chapter 5.) So the legend goes, Professor Bumpus was walking across the campus of Brown University, the day after a severe winter storm, and came across a number of motionless house sparrows on the ground. Bumpus collected the birds and brought them to his lab. Seventy-two birds recovered; 64 did not (Table 5.6.1).

Table 5.6.1. *Bumpus data set, summarized by sex of birds.*

House sparrows	Lived	Died
Female	21	28
Male	51	36

Bumpus reported differences in body size that correlated with survival (Bumpus 1899), and this report is often taken as an example of Natural Selection (cf. Johnston et al 1972). The Bumpus dataset is clearly a case of convenience sampling. It's also a case study: a report of a single incident. But given that is is a large sample $(n = 136)$, it is tempting to use the data to inform about about possible characteristics of the birds that survived compared with those that perished.

Another way we collect samples from populations is best termed haphazard. In graduate school I got the opportunity to study locomotor performance of whiptail lizards (*Aspidoscelis tigris, A. marmoratus, genus formerly Cnemidophorus***) across a hybrid zone in the Southwest United States (Dohm et al. 1998). During the day we would walk in areas where the lizards were known to occur and capture any individual we saw by hand. (This would sometimes mean sticking our hands down into burrows, which was always exciting — you never really knew if you were going to find your lizard or if you were going to find a scorpion, venomous

spider, or …) Lizards collected were returned to the lab for subsequent measures. Clearly, this was not **convenience sampling**; it involved a lot of work under the hot sun. But just as clearly, we could only catch what we could see and even the best of us would occasionally lose a lizard that had been spotted. Moreover, one suspects we missed many lizards that were present, but not in our view. Lizards that were underground at the time we visited a spot would not be seen nor captured by us; individual lizards that were especially wary of people (Bulova 1994) would also escape us. In other words, we caught the lizards that were catchable and could can only assume that they were representative of all of these lizards. Applying a grid or quadrant system to the area and then randomly visiting plots within the grid or quadrant would help, but still would not eliminate the potential for biased sampling we faced in this study.

Quota sampling implies selection of subjects by some specific criteria, weighted by the proportions represented in the population. It's different from stratified sampling because there is no random selection scheme: subjects are selected to be part of the study based on matching some criterion, and collection for that group stops when the sample number matches the proportion in the population. Consider our vitamin supplement survey. If the student population at Chaminade University was the reference population, and we have enough money to survey 100 students, then we would want a sample of 70 female students and 30 male students, representing the proportions of the student population.

Snowball sampling implies that you rely on word-of-mouth to complete sampling. After initial recruitment of subjects, sample size for the study increases because early participants refer others to the researchers. This can be a powerful tool for reaching underrepresented communities (e.g., Valerio et al 2016).

Types of Probability sampling

Random sampling is an example of probability sampling. As we defined earlier, simple random sampling requires that you know how many subjects are in the population (N) and then each subject has an equal chance of being selected: $p = \frac{1}{N}$

Examples of nonprobability sampling include:

- convenience sampling
- volunteer sampling
- judgement sampling

Convenience sampling (the first 20 people you meet at the library lanai); **volunteer sampling** (you stand in front of a room of strangers and ask for any ten people to come forward and take your survey — or more seriously, persons with a terminal disease calling a clinic reportedly known to cure the disease with a radical new, experimental treatment), and judgement sampling (to study tastes in fashion, you decide that only persons over six feet tall should be included because ….).

"Random" in statistics has a very important, strict meaning. As opposed to our day-to-day usage, random sampling from a population means that the *probability that any one individual is chosen to be included in a sample is equal*. Formally, this is called simple random sampling to distinguish it from more complex schemes. For a sampling procedure to be random requires a formal procedure for sampling a population with known size N).

For example, at the end of the semester, I may select the order for your talks at random. Thus, groups of students in this room are considered the population (groups of students are my sample unit, not individual students!). What is the probability that your group will be called first? Second? We need to know how many groups there are to conduct simple random sampling. Let's take an extreme and say that all groups have a size of one; there are 26 students in this room, so $p = \frac{1}{26}$ of being selected first.

Now to determine the probability of your group being selected second, we need to distinguish between two kinds of sampling:

- Sampling with replacement — after I select the first group, the first group is returned to the pool of groups that have not been selected. In other words, with replacement, your group could be selected first and selected second! The probability of being selected second then remains $\frac{1}{26}$
- Sampling without replacement — after I select the first group, then I have $26 - 1$ groups left to select the second group, so probability that your group will be second is $\frac{1}{25}$. The first group has already been selected and is not available, and so on.

Random sampling refers to how subjects are selected from the target reference population. **Random assignment**, however, describes the process by which subjects are assigned to treatment groups of an experiment. Random sampling applies to the **external validity** of the experiment: to the extent that a truly random sample was drawn, then results may be generalized to the

study population. Random assignment of subjects to treatment, however, makes the experiment **internally valid**: results from the experiment may be interpreted in terms of causality.

Additional sampling schemes

Simple random sampling is not the only option, but in many cases it is the most desirable. Consider our multivitamin study again. Perhaps studying the entire USA population is a bit extreme. How about working from a list of AARP members, sending out questionnaires to millions on this list, getting back about 20% of the questionnaires, sorting through the responses and identifying the respondent to diet categories? The researchers had nearly 500,000 persons willing and able to participate in their prospective study (Dawsey et al 2014). It's an enormous study. But is this really much better than our described lizard experiment? Let's count the ways: not all older people are members of the AARP (that 500,000? That's less than 1% of the 50 and older persons in the USA); a large majority of AARP members did not return surveys; some fraction of the returned surveys were not usable; how representative of diverse aged populations in the United States is AARP?

Simple random sampling may not be practical, particularly if sub-populations are present and members of the different sub-populations are not available to the researcher in the same numbers. Thus, samples are drawn in such a way as to represent the frequency within each sub-population. For a simple example, researchers conducting a controlled breeding program of mice don't use simple random sampling to choose pairs of mates; after all, random sampling without regard to sex of the mice would lead to some pairings of males only, or females only. Thus, the breeding strategy is to random sample from female mice and from male mice, and the stratification is sex of the mouse. Alternatively, breeders may select mice to form breeding pairs systematically: From a large colony with dozens of cages, the breeder may select one mouse from every third cage.

Stratified Random Sampling: Divide the reference population into groups, as many as needed. Then choose a simple random sample from each group. Combine those into the overall sample. For example, when I wanted a random sample of mice for my work, I called the supplier and requested that a total of 100 male and female mice be randomly selected from the five colonies they maintained. The reference population is the entire supply of mice at that company (at that time), but I wished to make sure that I got unrelated mice, so I needed to divide the population into groups (the five colonies) before my sample was constructed. Note that the size of the population must be known in advance, just like in simplified random sampling. In a more interesting example, the Social Security Administration conducts surveys of popular baby names by year. They post the top ten most popular names based on 1% or 5% (first strata), then by male/female (second strata).

Cluster Sampling: In many situations, the population is far too large or too dispersed and scattered for a list of the entire population to be known. And, a random approach ignores that there is a natural grouping — people live next to each other, so there are going to be things in common. A multi-stage approach to sampling will be better than simply taking a random sample approach. Most surveys of opinion (when conducted reputably) use a multi-stage method. For example, if a senator wishes to poll his constituents about an issue, his pollster will randomly select a few of the counties from his state (first stage), then randomly select among towns or cities (second stage), to obtain a list of 1000 people to call. In some instances, they might use even more stages. At each stage, they might do a stratified random sample on sex, race, income level, or any other useful variable on which they could get information before sampling. If you are interested in this kind of work, for starters see Couper and Miller (2009).

There are more types of sampling, and entire books written about the best way to conduct sampling. One important thing to keep in mind is that as long as the sample is large relative to the size of the population, each of the above methods generally will get the same answers (= the statistics generated from the samples will be representative of the population).

As long as some attempt is made to randomize, then you can say that the procedure is probability sampling. Nonprobability, or haphazard sampling, describes the other possibility, that is, each element is selected arbitrarily by a non-formal selecting of individuals… all the fish or birds that you catch may not be a random sample of those present in a population. For example,

- wild Pacific salmon do not feed on the surface, hatchery salmon feed on the surface.
- all the individuals who respond to a survey. Phone surveys, web surveys, person-on-the-street surveys… how random, how representative are they?

Sampling with computers

Sampling is usually easiest if a computer is used. Computers use algorithms to generate **pseudo-random numbers**. We call the resulting numbers pseudo-random to distinguish them from truly random physical processes (e.g., radioactive decay). For more information about random numbers, please see www.random.org.

If all you wish to do is select a few observations or you need to use a random procedure to select subjects prior to observations, then these websites can provide a very quick, useful tool.

Sampling in Microsoft Excel or LibreOffice Calc

Microsoft Excel is pretty good at sampling, but requires knowledge of included functions. Here are the steps to generate random numbers and select with and without replacement in Excel. I'll give you two cases.

1. For random numbers, enter the function **=rand()** in a cell, then drag the cell handle to fill in cells to N (in our case $N = 26$, so A1 to A26). This function generates a random (more or less!) number between 0 and 1. We want digits between 1 and 26, not fractions between 0 and 1, so combine **INT** function with RAND function:

=INT(27*RAND())

Note: To get between 1 and 9, multiply by 10 instead of 27; to get between 1 and 100, multiply by 101, etc.

In Excel, to sample with replacement, simply pick the first two cells (the algorithm Excel used already has conducted sample with replacement. See next item for method to sample without replacement in Excel. You have to have installed the Data Analysis Tool Pak. Here's instructions for Office 2010.)

If you have a Mac and Office 2008, there is no Data Analysis Tool Pak, so to get this function in your Excel, install a third-party add-in program (e.g., StatPlus, a free add-in, really nice, adds a lot of function to your Excel). If you have the 2011 version of Office for Mac, then the Data Analysis Tool Pak is included, but like your Windows counterparts, you have to install it (click here for instructions).

2. Let's say that we have already given each group a number between 1 and 26 and we enter those numbers in sequence in column A.

To sample without replacement, select Tools → Data Analysis… (if this option is not available, you'll have to add it — see Excel help for instructions, Fig. 5.6.1).

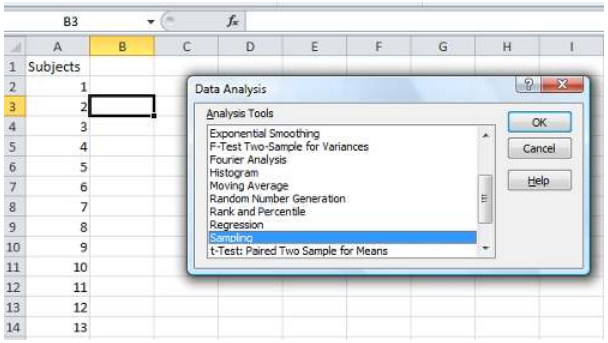


Figure 5.6.1: Screenshot of Sampling tool in Data Analysis menu, Microsoft Excel.

Enter the cells with the numbers you wish to select from. In our example, column `A` has the numbers 1 through 26 representing each group in our class. I entered `A:A` as the Input Range.

Next, select "Random" and enter the number of samples. I want two.

Click OK and the output will be placed into cell `B3` (my choice); I could have just as easily had Excel put the answer into a new worksheet.

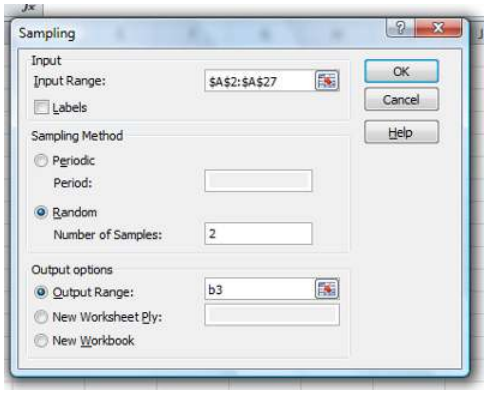


Figure 5.6.2: Screenshot of input required for Sampling in Data Analysis menu, Microsoft Excel.

To sample with replacement from column `A` (with our 1 through 26), type in the formula `B1`

```
=INT(27*RAND())
```

and the formula C1

```
=INDEX(A:A,RANK(B1,B:B))
```

then drag the cell handles to fill in the columns (first column B, then column C).

That'll do it for MS Excel or LibreOffice Calc.

Sampling with R (Rcmdr)

It's much easier to get samples with more control in Rcmdr (R) than in Excel. Sampling in R is based on the function called `sample()` and `sample.int()` . I will present just the `sample()` command here.

```
sample(x, size, replace = FALSE, prob = NULL)
```

For example, you want to sample ten integers between 1 and 10:

```
sample(10)
```

R output:

```
sample(10)
[1]  5  1 10  8  7  3  4  9  2  6
```

You have a list of subjects, A1 through A10:

```
subjects = c("A1","A2","A3","A4","A5","A6","A7","A8","A9","A10")
sample(subjects, 3, replace = FALSE, prob = NULL)
```

R output:

```
sample(subjects, 3, replace = FALSE, prob = NULL)
[1] "A5" "A2" "A9"
```

YOu could use this to arrange a random order for ten subjects:

```
sample(subjects, 10, replace = FALSE, prob = NULL)
[1] "A9" "A3" "A8" "A2" "A1" "A4" "A10" "A7" "A5" "A6"
```

Now try sampling with replacement. To do so, type in `TRUE` after `replace` in the `sample()` function. The R output follows:

```
sample(x, 10, replace = TRUE, prob = NULL)
[1]  5  5  3  5  3  7  3  5 10  6
```

R's randomness is based on psedu\nonumbers and is, therefore, not truly random (actually, this is true of just about all computer-based algorithms unless they are based on some chaotic process). We can use this pseudo part to our advantage: if we want to reproduce our "random" process, we can seed the random number algorithm to a value (e.g., 100), with the command in the Script Window:

```
set.seed(100)
```

For 10 random integers (e.g., observations), type in the Script window:

```
sample(10)
```

R returns the following in the Output Window:

```
sample(10)
[1] 4 5 7 3 10 9 2 1 6 8
```

Sampling was done without replacement.

Here's another selection round, first without setting a seed value:

```
sample(10)
[1] 10 4 7 3 8 1 2 9 6 5
```

Now, we try again to see if we get the same sample:

```
sample(10)
[1]  1  4  8  2  9  6  5 10  3  7
```

Now to demonstrate how setting the seed allows you to draw repeated samples that are the same. Note that I need to precede the sample command with a `set.seed()` call — when I do that, then the sampling is repeatable.

```
set.seed(100)
sample(10)
[1] 4 3 5 1 9 6 10 2 8 7
```

and try again

```
set.seed(100)
sample(10)
[1] 4 3 5 1 9 6 10 2 8 7
```

Additional R packages that help with sampling schemes include `sampling()` and `spatialsample`, which is part of the `BiodiversityR` package, which is available as a plugin for R Commander.

Questions

1. For our two descriptions of experiments in section 5.1 (the sample of patients; the sample of frogs), which sampling technique was used?
2. What purpose is served by `set.seed()` in a sampling trial?
3. True or False. If sample with replacement is used, a subject may be included more than once.
4. Use `sample()` with and without replacement on the object to create:

```
fruit <- c("apple", "banana", "grape", "kiwi", "pear", "pineapple", "tomato")
```

   a) set of 3

   b) set of 4

5. Consider our question, Does taking a multivitamin daily improve health? Imagine you have a grant willing to support a long-term prospective study to follow up to one thousand people for ten years. List at least three concerns with proposed solutions about how sampling of subjects for the study.
6. Imagine you wish to conduct a detailed survey to learn about student preferences. Your survey will include many questions, so you decide to ask just ten students. Student population is 70% female, 30% male.
    1. Assuming you select at random (simple random sampling), what is the chance that no male students will be included in your survey?
    2. You are able to increase the number of surveys to 20, 30, 40, or 50. What is the chance that no male students will be included in your survey for each of these increased sample numbers?
        1. What can you conclude about the effects of increasing survey sample size on representativeness of students for the survey?
7. Discuss how you could apply a stratified sampling scheme to this survey and whether or not this approach improves representativeness.
8. Why are random numbers generated by a computer called pseudorandom numbers?

5.7: Chapter 5 References

Alvarez, J. A., Garten, K. M., & Cook, D. G. (2021). Limb malformation in a foothill yellow-legged frog (*Rana boylii*) from Sonoma county, California. *Northwestern Naturalist*, *102*(3), 258-260.

Beck, D. D., Dohm, M. R., Garland, T., Ramírez-Bautista, A., & Lowe, C. H. (1995). Locomotor performance and activity energetics of helodermatid lizards. *Copeia*, 3, 575-585.

Benson, K., Hartz, A. J. (2000). A Comparison of Observational Studies and Randomized, Controlled Trials. *The New England Journal of Medicine* 342:1878-1886

Berger, R. L. (1990). Nazi science—the Dachau hypothermia experiments. *New England journal of medicine*, *322*(20), 1435-1440.

Bhatt A. (2010). Evolution of clinical research: a history before and beyond James Lind. *Perspectives in clinical research*, *1*(1), 6–10.

Binford, C. H. (1936). The history and study of leprosy in Hawaii. *Public Health Reports (1896-1970)*, 415-423.

Blainey, P., Krzywinski, M., & Altman, N. (2014). Points of significance: replication. *Nature Methods* 11, 879–880.

Bozkurt, B., Kamat, I., & Hotez, P. J. (2021). Myocarditis with COVID-19 mRNA vaccines. *Circulation*, *144*(6), 471-484.

Brandt, A. M. (1978). Racism and research: the case of the Tuskegee Syphilis Study. *Hastings center report*, 21-29.

Bulova, S. J. (1994). Ecological correlates of population and individual variation in antipredator behavior of two species of desert lizards. *Copeia*, 980-992.

Bumpus, H. C. 1898. Eleventh lecture. The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. (A fourth contribution to the study of variation.) Biological Lectures: Woods Hole Marine Biological Laboratory, 209-225. (link to volume at Google Books)

Concato, J., Shah, N., Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine* 342:1887-1892

Couper, M. P., & Miller, P. V. (2008). Web survey methods: Introduction. *Public Opinion Quarterly*, 72(5), 831-835.

Cox, K. L., Puddey, I. B., Morton, A. R., Burke, V., Beilin, L. J., McAleer, M. (1996). Exercise and weight control in sedentary overweight men: effects on clinic and ambulatory blood pressure. *Journal of Hypertension*14:779-790.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology*, 177(1), 7-11.

Darveau, C.-A., Hochachka, P. W., Welch, Jr., K. C., Roubik, D. W., Suarez, R. K. (2005). Allometric scaling of flight energetics in Panamanian orchid bees: a comparative phylogenetic approach. *Journal of Experimental Biology* 208:3581-3591

Dawsey, S. P., Hollenbeck, A., Schatzkin, A., & Abnet, C. C. (2014). A prospective study of vitamin and mineral supplement use and the risk of upper gastrointestinal cancers. *PLoS One*, 9(2), e88774.

Diez Roux, A. V. (2004). The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. *Epidemiologic reviews*, 26(1), 104-111.

Dohm, M. R., Garland Jr, T., Cole, C. J., & Townsend, C. R. (1998). Physiological variation and allometry in western whiptail lizards (*Cnemidophorus tigris*) from a transect across a persistent hybrid zone. *Copeia*, 1-13.

Dohm, M. R., Richardson, C. S., & Garland Jr, T. (1994). Exercise physiology of wild and random-bred laboratory house mice and their reciprocal hybrids. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 267(4), R1098-R1108.

Doll, R. (1998). Uncovering the effects of smoking: historical perspective. *Statistical Methods in Medical Research* 7:87-117

Earp, B. D. (2011). Can science tell us what's objectively true. *The New Collection*, *6*(1), 1-9.

Elwood, J. M. (2013). Commentary: On representativeness. *International Journal of Epidemiology* 42:1014-1015.

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? *Journal of the American Medical Association*, 283(20), 2701-2711.

Fang, L., Karakiulakis, G., & Roth, M. (2020). Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?. *The Lancet. Respiratory Medicine*, 8*(4), e21*.

Foster, K. A., Oster, C. G., Mayer, M. M., Avery, M. L., & Audus, K. L. (1998). Characterization of the A549 cell line as a type II pulmonary epithelial cell model for drug metabolism. *Experimental cell research*, 243(2), 359-366.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, 317:141-145.

Fritts, T. H., Rodda, G. H. (1998). The role of introduced species in the degradation of island ecosystems: A case history of Guam. *Annual Review of Ecology and Systematics* 29:113-140.

Gabriella, R. (2012) Studying drugs in all the wrong people. *Scientific American Mind* 23:34-41.

Goodman, M., LaKind, J. S., Fagliano, J. A., Lash, T. L., Wiemels, J. L., Winn, D. M., … & Mattison, D. R. (2014). Cancer cluster investigations: review of the past and proposals for the future. *International journal of environmental research and public health*, *11*(2), 1479-1499.

Hodge, F. S. (2012). No meaningful apology for American Indian unethical research abuses. *Ethics & Behavior*, *22*(6), 431-444.

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biology*, 13(7), e1002190.

Houle, D., Pélabon, C., Wagner, G. P., Hansen, T. F. (2011). Measurement and meaning in biology. *The Quarterly Review of Biology* 86:3-34

Howard, D. H., Kenline, C., Lazarus, H. M., LeMaistre, C. F., Maziarz, R. T., McCarthy Jr, P. L., … & Majhail, N. S. (2011). Abandonment of high-dose chemotherapy/hematopoietic cell transplants for breast cancer following negative trial results. *Health services research*, *46*(6pt1), 1762-1777.

Jennings, S., Reynolds, J. D., Mills, S. C. (1998). Life history correlates of responses to fisheries exploitation. *Proceedings of the Royal Society B: Biological Sciences* 265:333-339

Johnston, R. F., Niles, D. M., & Rohwer, S. A. (1972). Hermon Bumpus and natural selection in the house sparrow *Passer domesticus*. *Evolution*, (26):20-31.

Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *British Medical Journal*, 326(7404), 1453-1455.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., Altman, D. G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology* 8:e1000412

Lariviere, W. R., Chesler, E. J., Mogil, J. S. (2001). Transgenic Studies of Pain and Analgesia: Mutation or Background Genotype? *Journal of Pharmacology and experimental therapeutics* 297:467-473

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience* 11:5

Losos, J. B., Creer, D. A., & Schulte Ii, J. A. (2002). Cautionary comments on the measurement of maximum locomotor capabilities. *Journal of Zoology*, *258*(1), 57-61.

Luz, R. K., Martínez-Álvarez, R. M., De Pedro, N., & Delgado, M. J. (2008). Growth, food intake regulation and metabolic adaptations in goldfish (*Carassius auratus*) exposed to different salinities. *Aquaculture*, 276(1-4), 171-178.

Marshall, M., Ferguson, I. D., Lewis, P., Jaggi, P., Gagliardo, C., Collins, J. S., … & Guzman-Cottrill, J. A. (2021). Symptomatic acute myocarditis in seven adolescents following Pfizer-BioNTech COVID-19 vaccination. *Pediatrics*, 2.

McCambridge, J., Sorhaindo, A., Quirk, A., & Nanchahal, K. (2014). Patient preferences and performance bias in a weight loss trial with a usual care arm. *Patient education and counseling*, *95*(2), 243-247.

Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota Studies in the philosophy of science: Vol. IV. Analyses of theories and methods of physics and psychology* (pp. 373-402). Minneapolis: University of Minnesota Press.

Medical Research Council Investigation (1948). STREPTOMYCIN treatment of pulmonary tuberculosis. *British Medical Journal*, *2*(4582), 769–782.

Milham, S. (2004). A cluster of male breast cancer in office workers. *American Journal of Industrial Medicine* 46:86-87.

Nelson, M. R., et al. (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *American Journal of Human Genetics* 83:347-358

Nethery, R. C., Yang, Y., Brown, A. J., & Dominici, F. (2020). A causal inference framework for cancer cluster investigations using publicly available data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 1253-1272.

Nijhawan, L. P.; Janodia, M. D.; Muddukrishna, B. S.; Bhat, K. M.; Bairy, K. L.; Udupa, N. & Musmade, P. B. (2013) Informed consent: Issues and challenges. *Journal of Advanced Pharmaceutical Technology & Research* 4:134-140

O'Leary, S. T., & Maldonado, Y. A. (2021). Myocarditis After SARS-CoV-2 Vaccination: True, True, and… Related?. *Pediatrics*, *148*(3).

Omair A. (2015) Selecting the appropriate study design for your research: Descriptive study designs. *Journal of Health Specialties* 3:153-156.

Omair A. (2016) Selecting the appropriate study design: Case–control and cohort study designs. *Journal of Health Specialties* 4:37-41.

Pannucci, C. J. & Wilkins, E. G. (2010) Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery* 126, 619-625

Pool, R. (1990). Is there an EMF-cancer connection? *Science*, 249(4973), 1096-1099.

Reader, R. J. (1992). Herbivory as a confounding factor in an experiment measuring competition among plants. *Ecology* 73(1):373-376

Roelcke, V. (2004). Nazi medicine and research on human beings. *The Lancet*, *364*, 6-7.

Rothman, K. J., Gallacher, J. E. J., Hatch, E. E. (2013). Why representativeness should be avoided. *International Journal of Epidemiology* 42:1012-1014.

Rothwell, E., Brassil, D., Barton-Baxter, M., Brownley, K. A., Dickert, N. W., Ford, D. E., … & Wilfond, B. S. (2021). Informed consent: Old and new challenges in the context of the COVID-19 pandemic. *Journal of Clinical and Translational Science*, *5*(1).

Scott, A. F. (1999). Malformed southern leopard frogs (*Rana sphenocephala utricularius*). *Journal of the Tennessee Academy of Science*, *74*(3-4), 61-63.

Shuster, E. (1997). Fifty Years Later: The Significance of the Nuremberg Code. *The New England Journal of Medicine* 337: 1436-1440.**

Sigmund, C. D. (2000). Viewpoint: Are studies in genetically altered mice out of control? *Arteriosclerosis, Thrombosis, and Vascular Biology* 20:1425-1429

Simons, D. J. (2014). The value of direct replication. *Perspectives on psychological science*, *9*(1), 76-80.

Stephens, P. A., Buskirk, S. W., del Rio, C. M. (2006). Inference in ecology and evolution. *Trends in Ecology & Evolution* 22:192-197

Stevens, S. S. (1946) On the Theory of Scales of Measurement. *Science* 103:677-680

Suba, E. J. US-funded measurements of cervical cancer death rates in India: scientific and ethical concerns. *Indian Journal of Medical Ethics*11:167-175.

Temple, R.& Ellenberg, S. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Annals of Internal Medicine* 133: 455-463

Tension headache symptoms. Internet, cited 2016 Apr. 20. Retrieved from: http://www.mayoclinic.org/diseases-conditions/tension-headache/basics/symptoms/con-20014295

Thomas, M., & Bomar, P. A. (2018). Upper Respiratory Tract Infection. In StatPearls Internet. StatPearls Publishing. At https://www.ncbi.nlm.nih.gov/books/NBK532961/

Thompson, C. B., & Panacek, E. A. (2006). Research study designs: experimental and quasi-experimental. *Air medical journal*, *25*(6), 242-246.

Thompson, S. J., Auslander, W. F., White, N. H. (2001). Comparison of Single-Mother and Two-Parent Families on Metabolic Control of Children With Diabetes. *Diabetes Care* 24(2): 234-238.

Tignanelli, C. J., Ingraham, N. E., Sparks, M. A., Reilkoff, R., Bezdicek, T., Benson, B., … & Puskarich, M. A. (2020). Antihypertensive drugs and risk of COVID-19?. *The Lancet Respiratory Medicine*. 8:e30-e31.

Treece, J. W. Jr (1990, retrieved January 9 2023). Daniel and the Classic Experimental Design. *Institute and Creation Research*, https://www.icr.org/article/daniel-classic-experimental-design/.

Tyrer, S., & Heyman, B. (2016). Sampling in epidemiological research: issues, hazards and pitfalls. *BJ Psych Bulletin*, 40(2), 57-60.

Valerio, M. A., Rodriguez, N., Winkler, P., Lopez, J., Dennison, M., Liang, Y., & Turner, B. J. (2016). Comparing two sampling methods to engage hard-to-reach communities in research priority setting. *BMC Medical Research Methodology*, 16(1), 146.

Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO reports*, *13*(4), 291-296.

Wacholder, S., McLaughlin, J. K., Silverman, D. T., & Mandel, J. S. (1992). Selection of controls in case-control studies: I. Principles. *American journal of epidemiology*, *135*(9), 1019-1028.

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992). Selection of controls in case-control studies: II. Types of controls. *American journal of epidemiology*, *135*(9), 1029-1041.

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992). Selection of controls in case-control studies: III. Design options. *American journal of epidemiology*, *135*(9), 1042-1050.

Walker, G. (2009). Beyond distribution and proximity: exploring the multiple spatialities of environmental justice. *Antipode*, 41(4), 614-636.

Wang, Y., Zhang, D., Du, G., Du, R., Zhao, J., Jin, Y., … & Hu, Y. (2020). Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*.

Whitley, E., Ball, J. (2002). Statistics review 2: Samples and populations. *Critical Care* 6:43-48.

Zeilinger, J., Steger-Hartmann, T., Maser, E., Goller, S., Vonk, R., & Länge, R. (2009). Effects of synthetic gestagens on fish reproduction. *Environmental toxicology and chemistry*, *28*(12), 2663-2670.

Zoorob, R., Sidani, M. A., Fremont, R. D., & Kihlberg, C. (2012). Antibiotic use in acute upper respiratory tract infections. *American family physician*, *86*(9), 817-822.

CHAPTER OVERVIEW

6: Probability and Distributions

Introduction

Probability is how likely the occurrence of some event is. Thus, an important concept to appreciate is that in many cases, like R.A. Fisher's Lady tasting tea analogy, we can count in advance all possible outcomes of an experiment. On the other hand, for many more experiments, we cannot count all possible outcomes of the **sample space**, either because they are too numerous or simply unknowable. In such cases, applying **theoretical probability distributions** allow us to circumvent the countability problem. Whereas **empirical probability** distributions are frequency counts of observations, theoretical probabilities are based on mathematical formulas.

Much of classical inferential statistics, especially the kind one finds in introductory courses like ours, are built on **probability distributions**. ANOVA, t-tests, linear regression, etc., are **parametric tests** and assume errors are distributed according to a particular type of distribution, the **normal** or **Gaussian distribution**.

A probability distribution is a list of probabilities for each possible outcome of a **discrete random variable** in an entire population. Depending on the data type, there are many classes of probability distributions. In contrast, probability density functions are used to for **continuous random variables**. This chapter begins with basics of probability, then gently introduces probability distributions. In the other sections of this chapter we describe several probability density functions. Emphasis is placed on the normal distribution, which underlies most parametric statistics.

6.1: Some preliminaries

Introduction

OK, you say, I get it: statistics is important, and if I am to go on as a biologist, I should learn some biostatistics. Let's get on with it, start with the equations and the problems already!

Before we review **probability theory** and introduce **risk analysis** I want to spend some time to emphasize that at issue is critical thinking, so please bear with me (Fig. 6.1.1).

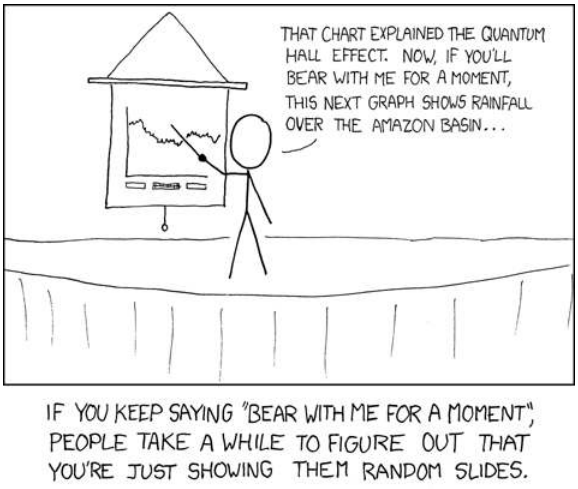


Figure 6.1.1: https://xkcd.com/365/

How likely?

As we start our journey in earnest, we start with the foundations of statistics, **probability**. Probability is something we measure, or estimate, about whether something, an **event**, will occur. We speak about *how likely* an event is to occur, and this is quantified by the probability. Probabilities are given a value between 0 and 1: at 0% chance, the event will not happen; at 100% chance, the event is certain to happen.

If probability is the chance that an event will occur, then **risk** is the probability of an event occurring over a specified period of time. I introduce our discussion of probability through a risk analysis. I like to start this discussion by relaying something I overheard, while we were all standing on a lava field on the slopes of Kilauea back in November of 1998 (Kilauea erupted with lava flow more or less continually between 1983 and 2018; on 5 Jan 2023, it started up again).

Figure 6.1.2: View of Kamokuna Lava Bench, eruption of Pu`u `O`o, Kilauea, November 1998. Photo by S. Dohm.

The Volcanoes National Park hadn't established barricades at the end of Chain of Craters Road, and people were walking to see new lava flows. The night we went, we met a park ranger who announced to us that the Park Service believed it was unsafe for us to walk out to see new lava flows because the area was unstable. Someone (not in my group), snapped back, "Oh, what are the chances that that will happen?" Of course, the ranger couldn't quote a chance between zero and 100% for that particular evening. The ranger was saying the risk had increased, based on their subjective, but experienced, opinion.

As you can see in Figure 6.1.2, we went anyway. I have been thinking about her question ever since. We were lucky — some of the same area collapsed two weeks later (USGS update 16 December 1998).

Cool picture though.

Multiple events

I have two coins, a dime and a quarter, in my pocket; when I place the coins on the table, what are the chances that both coins will show heads? A blood sample from a crime scene was typed for two **Combined DNA Index System** (CODIS) Short Tandem Repeat (STR) loci, THO1 (allele 9.3) and TPOX (allele 8), the same allele types for the defendant. What are the chances of a random match, that someone other than the defendant has the same genetic profile? For two or more **independent events** we can get the answers by using the **product rule**.

$$P\left(H_{\text{dime}} \text{ and } H_{\text{quarter}}\right) = P\left(H_{\text{dime}}\right) \times P\left(H_{\text{quarter}}\right) = 0.5 \times 0.5 = 0.25$$

The two coins are independent; therefore, the chance that both are placed heads up is 25% — we would expect to see this **combined event** one out of every four times. This is an illustration of the **counting rule**, aka **fundamental counting principle**: if there are n ways to do one thing (n elements in set A), and m ways to do another thing (m elements in set B), then there are $n \cdot m$ ways to do both things (combination of elements of A and B sets).

For the DNA profile CODIS problem (cf. Chapter 4, National Research Council 1996), the two alleles are both the most common observed in US Caucasian population at 30.45% and 54.7%, respectively (Moretti et al 2016). Assuming the individual was homozygous at both loci (i.e., $\text{THO1}_{9.3,9.3}$ and $\text{TPOX}_{8,8}$), then the genotype frequencies (p^2) are:

$\text{THO1}_{9.3,9.3} = 0.3045^2 = 0.093$

$\text{TPOX}_{8,8} = 0.547^2 = 0.299$

Since THO1 is located on the p-arm of chromosome 11, and TPOX is on the p-arm of chromosome 2, the two loci are independent and therefore should be in linkage disequilibrium. We can use the product rule to get the probability of the DNA profile for the sample, 2.8%:

$$P\left(\text{THOI and TPOX}\right) = P(\text{THOI}) \times P(\text{TPOX}) = 0.093 \times 0.299 = 0.028$$

If two events are not independent, then the product rule cannot be used. For example, CODIS STR D5S818 and CSF1PO are both located on the q-arm of chromosome 5 and are therefore linked and not independent (the recombination frequency is about 0.25). The common allele for D5S818 is 11 at 40.84% and for CSF1PO the allele is 12 at 34.16%. Thus the chance of gettting the two most common alleles is not simply the product rule result of 14%; instead, we need to view this problem as one of dependent events.

Kinds of probability

So, how does one go about estimating the likelihood that a particular event will occur, whether it is the collapse of a lava delta, or that a person will have a heart attack? The probability of lava delta collapse or of heart attack are examples of **empirical probability**, as opposed to a **theoretical probability**. Despite many years of effort, we have no applicable theory that we can apply to say, if a person does this, and that, then a heart attack will happen. But we do have a body of work documenting how often heart attacks occur, and when they occur in association with certain risk factors. Similarly, progress is being made to determine markers of risk of lava field collapse (Di Traglia et al. 2018). Analogously, this is the essential goal of risk analysis in epidemiology. We know of associations between cholesterol and heart attack risk, for example, but we also know that high cholesterol does not raise the probability of the event (heart attack) to 100%. How is this **uncertainty** part of statistics? Or perhaps you are a molecular scientist in training and have learned about how to assess results of a Western blot where typically the results are scored as “yes” or “no.” How is this relevant to statistics and probability?

A misconception about statistics and statistical thinking

There’s a long history of skepticism of conclusions from health studies, in part because it seems the advice flips. For example,

- Coffee is bad for you (Medical News Today January 2008)
- Coffee is good for you (NBC News July 2018)
- Even light drinking can be harmful to health (Science Daily, January 2022)
- Seven science-backed reasons beer is good for you (NBC News August 2017)

- Meat and cheese may be as bad for you as smoking (Science Daily, March 2014)
- Cheese actually isn't bad for you (WIRED, February 2021)

The common thread is these studies are assessment of risk: about studies that seem to conclude only with statements of probability.

> **Note:**
>
> This "flipping" seems as much a function of reporting bias — the studies are not directly comparable — and may just be clickbait.

Perhaps you may have heard …? "*There are lies and then there is statistics*". The full quote reads as follows:

> Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: "There are three kinds of lies: lies, damned lies and statistics." – *Autobiography of Mark Twain* (www.twainquotes.com/Lies.html).

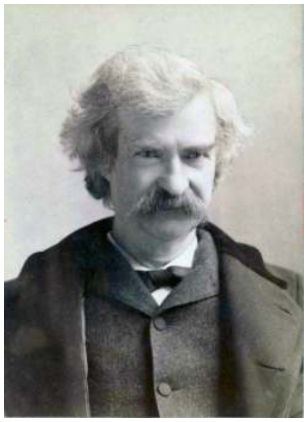


Figure 6.1.3: Mark Twain. Image from The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Photography Collection, The New York Public Library. "Mark Twain in Middle Life" The New York Public Library Digital Collections. 1860 – 1920. https://digitalcollections.nypl.org/items/510d47d9-baec-a3d9-e040-e00a18064a99

Twain attributed the remarks to Benjamin Disraeli, the prime minister of Britain during much of Queen Victoria's reign, but others have not been able to document this utterance to that effect. Still others believe that the "3-lies" quote belongs to Leonard Courtney, an English mathematician and statistician (1847-1929) (see University of York web site for more).

What is meant by this quote? Tossing aside cynicism — or healthy skepticism of authority that one should have, whether that authority is a scientist or a politician (within reason, please!) — what this quote means is that it seems that results of very similar studies are in conflict. To some, this is part of the **replication crisis** in science (Baker 2016). There's a perception that one can say just about anything with a number. Partly this is a matter of semantics, but also there is legitimacy to this concern. However, it is not necessarily the case that statistics have been intentionally done to mislead; rather, there is evidence that researchers are not always using proper statistical procedures.

One word, several meanings

We use the term statistics in multiple ways, all correct, but not all equal. For example, a statistic may refer to a number used to describe a population characteristic. From the 2010 U.S. census, we learn that the racial (self-reported) make-up of the U.S. population (then at 303 million) was 72.4% "white" and 12.6% "black" self-reported. In this sense, a statistic is something you calculate as a description. Do you recall the distinction between "statistic" and "statistics" discussed in Chapter 2?

> **Note:**
>
> This confusion is not restricted to the province of the of beginners. For example, I stumbled upon another imputation of the "Lies, Damned Lies, and Statistics" in a header of a published article (Baker et al 2014) in which the authors argued that more than 50% of papers published over a two-year period on experimental autoimmune encephalomyelitis (EAE) in rodents applied the wrong statistical procedures. The disagreement in this case had to do with **data types**; the outcome variable for EAE should be **ordinal**, but as many as half the authors reportedly (according to Baker et al) proceeded to calculate means and

conduct parametric statistical tests. Medians and not means are appropriate descriptive statistics for ordinal data types. Data types and descriptive statistics were covered in Chapter 3: Exploring data.

Secondly, two studies essentially about the same topic, yet reaching seemingly different conclusions, may differ in the assumptions employed. It should be obvious that if different assumptions are used, researchers may reach different outcomes.

Finally, how we communicate statistics can be misleading. For example, use of percentages in particular can be confusing, especially in communication of the chance that some event may happen to us (e.g., incidence of disease, or number of new cases in a specific time period, compared to prevalence of a disease, or number of cases of a disease in a specific period of time). On the one hand, percentages seem easy. A percentage is simply a proportion multiplied by 100%, and takes any value between 0% and 100%.

When a product says that it kills 99.99% of all germs on contact, do you feel better? Here's a cartoon to consider as you think about that statement.

Figure 6.1.4: xkcd comic strip, from https://imgs.xkcd.com/comics/hand_sanitizer.png

Of course numbers cannot be used to justify simultaneously mutually exclusive conclusions, but being able to recognize careless (or deliberate) miscommunication with numbers, well, this needs to be part of your skill set. As you read this next section, I ask that you consider:

- are the correct statistical descriptors in use?
- what assumptions are being made?
- does the reporting of percentages lead to clear conclusions?

Some concluding thoughts about "lies and statistics"

Statistics is tricky because there are assumptions to be made. And you have to be clear in your thinking.

If the assumptions hold true, then we aren't lying, and Twain had it wrong.

But if we disagree on the assumptions, then we will necessarily have to disagree on the conclusions drawn from the calculated numbers (the statistics). Risk analysis in particular, but statistics in general, is a tricky business because many assumptions need to be made, and we won't necessarily have all of the relevant information available to make sure our assumptions are truthful. But it is the assumptions that matter: if we agree with the assumptions that are made, then we have confidence in the conclusions drawn from the statistics.

In a typical statistics course, we would spend a bunch of time on probability. We will here as well, but in the context of risk analysis and in the other contexts, in a less than formal presentation on the subject of probability. For example, in talking about inference, the testing of null hypotheses and estimating the probability that the null hypothesis is correct, I will say things like, "Imagine we repeated the experiment a million times — how many times by chance would we think a correct null hypothesis would nonetheless be rejected?"

There's no real substitute for a formal course in probability theory, and you should be aware that this foundation is pretty important if you go forward with biostatistics and epidemiology. For now, I will simply refer you to chapters 1, 2 and 3 of a really nice online book on probability from one of the masters, Richard Jeffrey (1926 – 2002; click here to go to Wikipedia). Much of what I will present to you follows from similar discussions.

My aim is to teach you what you need about probability theory by the doing. In the next couple of days we will deal with an aspect of risk analysis, namely a consideration of CONDITIONAL PROBABILITY, and Baye's Theorem that will help you evaluate claims such as the one made for airline safety. Risk analysis is tricky, but it is not a subject above and beyond our abilities; by applying some of the rules of statistical reasoning, we can check claims based on statistics. A healthy degree of skepticism is part of becoming a scientist. Do try this at home!

Some examples to consider: what is the relative risk for the following scenarios?

1. Drug testing at the workplace: risk of a worker who does not use illegal drugs registering positive (false positives in drug testing);

2. Positive HIV from blood sample from USA male with no associated risks (e.g., intravenous drug user), false positives in HIV testing; false positives with mammography;

3. Benefits versus risks of taking a statin drug (drug that reduces serum cholesterol levels) to a person with no history of heart disease;

4. Is it safer to travel by car or by airliner? We'll break this problem down in the next section.

What we are looking for is the probability of an occurrence of a particular event, e.g., that a person who does not use illegal drugs may nonetheless test positive; we are looking for a way to make rational decisions and understanding probability is the foundation.

Questions

1. What do you make of the claim (joke) that "*There are lies and then there are statistics*?"
2. For the various proportions listed, can these also be considered to be rates?
3. Distinguish between empirical and theoretical probability; use examples.
4. CODIS STR D5S818 and CSF1PO are both located on the q-arm of chromosome 5, and are therefore linked and not independent (the recombination frequency is about 0.25). The common allele for D5S818 is 11 at 40.84% and for CSF1PO the allele is 12 at 34.16%. Given that the person has allele 11 at D5S818 (genotype 11,11), what are the chances that they also have allele 12 at CSF1PO (genotype 12,12)?

6.2: Ratios and probabilities

Introduction

Let's define our terms. An **event** is some occurrence. As you know, a ratio is one number, the numerator, divided by another, called the denominator. A **proportion** is a ratio where the numerator is a part of the whole. A **rate** is a ratio of the frequency of an event during a certain period of time. A rate may or may not be a proportion, and a ratio need not be a proportion, but proportions and rates are all kinds of ratios. If we combine ratios, proportions, and/or rates, we construct an **index**.

Ratios

Yes, data analysis can be complicated, but we start with this basic idea. Much of the statistics is based on frequency measures, e.g., ratios, rates, proportions, indexes, and scales.

Ratios are the association between two numbers, one random variable divided by another. Ratios are used as descriptors and the numerator and denominator do not need to be of the same kind. Business and economics are full of ratios. For example, Return on investment (ROI) equals net income divided by number of shares outstanding, the Price-Earnings, or P/E ratio, is the ratio of the price of a stock to the earnings per stock, as well as many others are used to summarize performance of a business, and to compare performance of one business against another. Ratios are a deceptively convenient way to standardize a variable for comparisons, i.e., how many times one number contains another. For example, when estimating bird counts for different areas, or different birding effort (intensity, time searched), we may correct counts by accounting for area in which counts were made or the total time spent counting, for a per-unit ratio (Liermann et al 2004).

Practice: There were 1,326 day undergraduate students enrolled in 2014 at Chaminade University of Honolulu and the Sullivan Library added 8469 new items (ebooks, journals, etc.,) to its collection during 2014. What is the ratio of new items per student?

$$\frac{8469 \text{ items}}{1326 \text{ students}} = 6.39 \text{ items per student}$$

Data collected from Chaminade University website at www.chaminade.edu on 3 July 2014.

Practice: For another example, what is ratio of annual institutional aid a student at Chaminade University may expect to receive compared to a student at Hawaii Pacific University?

$$\frac{\text{aid}_{\text{Chaminade}}}{\text{aid}_{\text{HPU}}} = \frac{\$8491}{\$3897} = 2.3$$

Fold-change

To compare the ratio between two quantities, e.g., to compare mRNA expression levels of genes from organisms exposed to different conditions, researchers may report **fold-change**.

An example of calculation of fold change is rates of the expression from cells exposed to heavy metal divided by expression under basal conditions. Gene expression under different treatments may be evaluated by calculating fold-change as the log base 2 of the ratio of expression of a gene for one treatment divided by expression of the same gene from control conditions. Copper is an essential trace element, but excess exposure to copper is known to damage human health, including chronic obstructive pulmonary disease. One proposed mechanism is that cell injury promotes an epithelial-to-mesenchyme shift. In a pilot study we investigated gene expression changes by quantitative real-time polymerase chain reaction (qPCR) in a rat lung Type II alveolar cell line exposed to copper sulfate compared to unexposed cells. We recorded cycle threshold values, C_T, for each gene, where C_T is the number of cycles required for the fluorescent signal to exceed background levels; C_T is inversely proportional to amount of cDNA (mRNA) in the sample. Genes investigated were ECAD, FOXC2, NCAD, SMAD, SNAI1, TWIST, and VIM, with ATCB as reference gene. ECAD expression is marker of epithelial cells, whereas FOXC2, NCAD, SNAI1, TWIST, and VIM expression marker of mesenchymal cells. After calculating $2^{-\Delta\Delta C_T}$ values, geometric means of normalized values of three replicates each are shown in Table 6.2.1.

Note:

Logarithm transform is used because gene expression levels vary widely on the original scale and any log-transform will reduce the variability. log-base 2 is used for fold-change in particular because it is easy to interpret and provides symmetry (all log-transforms provide this symmetry). For example, `log(1/2, 2)` returns -1, while `log(2/1, 2)` returns $+1$.

Thus, when using base 2, we see a decrease by half or doubling of original scale is a fold change of ± 1. In contrast, `log(1/2, 10)` returns -0.301, while `log(2/1, 10)` returns $+0.301$.

Table 6.2.1. Mean $2^{-\Delta\Delta C_T}$ and fold change of gene expression values from qPCR for several genes from a rat lung cell line.

	Control	Copper-sulfate	Fold change
ECAD	34.6	35.7	0.6
NCAD	28.5	24.0	27.2
SMAD	29.5	25.0	28.2
SNAI1	25.5	28.1	0.2
FOXC2	27.6	27.0	1.9
VIM	23.1	16.4	134.4
TWIST	25.1	22.9	5.6

At face value, there appears to be some evidence that following a four-hour exposure to copper sulfate in media, the epithelial cell line adopted gene expression profile of mesenchyme-like cells. However, the weakness of fold-change is clear from Table 6.2.1: the quantity is sensitive to small values. ECAD expression in the cell line is low, thus the treated cells go through high numbers of PCR cycles (mean = 36) and control cells not much fewer (mean = 34.4).

Note:

Calculation of $2^{-\Delta\Delta C_T}$ is included. Geomean C_T were

	Control	$CuSO_4$
ACTB	32.2	32.5
NCAD	28.5	24.0

For control cells, $\Delta C_{T\ control} = C_{T\ GO1} - C_{T\ Ref} = 28.5 - 32.2 = -3.7$

For treatment cells, $\Delta C_{T\ treatment} = C_{T\ GO1} - C_{T\ Ref} = 24.0 - 32.5 = -8.5$

and $\Delta\Delta C_T = \Delta C_{T\ treatment} - \Delta C_{T\ control} = -4.8$

$2^{-\Delta\Delta C_T} = 28$

Table value differs by rounding

Rates

Rates are a class of ratios in which the denominator is some measure of time. For example, the four year graduation rate of some Hawaii universities are shown in Table 6.2.2.

Table 6.2.2. Percent students graduation with bachelor's degrees within four years or six years (cohort 2014, data source NCES.ed.gov).

School	Private/Public	Four-year, Percent graduation	Six-year, Percent graduation
Chaminade University	Private (non-profit)	43	58
Hawaii Pacific University	Private (non-profit)	31	46
University of Hawaii – Hilo	Public	15	38
University of Hawaii – Manoa	Public	35	62

School	Private/Public	Four-year, Percent graduation	Six-year, Percent graduation
University of Hawaii – West Oahu	Public	16	39
University of Phoenix	Private (for profit)	0	19

Examples of rates

Rates are common in biology. To name just a few:

- Basal metabolic rate (BMR), often measured by indirect calorimetry, reported in units kilo Joules per hour.
- Birth and death rates, components of population growth rate.
- Phred quality score, error rates of incorrectly called nucleotide bases by the sequencer
- Growth rate, which may refer to growth of the individual (somatic growth rate), or increase of number of individuals in a population per unit time
- Molecular clock hypothesis, rate of amino acid (protein) or nucleotide (DNA) substitution is approximately constant per year over evolutionary time.

Proportions

Proportions are also ratios, but they are used to describe one part to the whole. For example, 902 women (self-reported) day undergraduate students enrolled in 2014 at Chaminade University in Honolulu, Hawaii.

Practice: Given that the total enrollment for Chaminade in 2014 was of 1,326, calculate the proportion of female students to the total student body.

$$\frac{902}{1326} = 0.68$$

Comparing proportions

In some cases you may wish to compare two proportions or two ratios. The hypothesis tested is the difference between the two ratios, and the test is if the confidence interval of the difference includes zero. If it does, then we would conclude there is no statistical difference between the two proportions. In R, use the `prop.test` function. For example, 63 women were on team sport rosters at Chaminade in 2014, a proportion of 59% of all student athletes $(n = 106)$. Recall from the example above that women were 68% of all students at Chaminade University. Title IX compliance requires that a university "maintain policies, practices and programs that do not discriminate against anyone on the basis of gender" (NCAA, http://www.ncaa.org/about/resources/inclusion/title-ix-frequently-asked-questions). In terms of athletic programs, then, universities are required to provide participation opportunities for women and men that are substantially proportionate to their respective rates of enrollment of full-time undergraduate students (NCAA, http://www.ncaa.org/about/resources/inclusion/title-ix-frequently-asked-questions)

Consider Chaminade University: Is there a statistical difference between proportion of women athletes and their proportion of total enrollment? We introduce statistical inference in Chapter 8, but for now, this is a test of the null hypothesis that the difference between the two proportions is zero.

At the R prompt type (remember, anything after the # sign is a comment and ignored by R).

```
women = c(62,902)  #where 62 is the number of women athletes and 902 is the number of
students = c(106,1326)  #106 is the number of student athletes and 1326 is all studen
prop.test(women,students)  #the default is a two-tailed test, i.e., no group differen
```

And R returns

```
2-sample test for equality of proportions with continuity correction
 data: women out of students
```

```
X-squared = 3.6331, df = 1, p-value = 0.05664
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.197532407 0.006861073
sample estimates:
 prop 1 prop 2
0.5849057 0.6802413
```

What is the conclusion of the test?

When you compare two groups, you're asking whether the two groups are equal (the null hypothesis). Mathematically, that's the same as saying the difference between the two groups is equal to zero.

First check the lower and upper limits of the confidence interval. A **confidence interval** is one way to report a range of plausible values for an estimate (see Ch 7.6 – Confidence intervals). It's called a confidence interval because a probability is assigned to the range of values; a 95% confidence interval is interpreted as we're 95% certain the true population value is somewhere between the reported limits. For our Chaminade University Title IX question, recall that we are asking whether the value of zero is included. The lower limit was -0.1975 and some change; the upper limit was 0.0068 and some change. Thus, zero is included and we would conclude that there was no statistical difference between the two proportions.

The second relevant output to look at is the **p-value**, or **probability value**. If the p-value is less than 5%, we typically reject the tested hypothesis. We will talk more about p-values and their relationship to inference testing in Chapter 8; for now, pay attention to the confidence interval (introduced in Chapter 3.4); if zero is included, then we conclude no substantial differences between the two proportions.

Indexes

Indexes are composite statistics that combine indicators. Indexes are common in business and economics, e.g., Dow Jones Industrial average combines stock prices from 30 companies listed on the New York Stock Exchange.

Some indexes presented in this book include

- Grade point average
- Body Mass Index (BMI)
- Comet assay indexes (tail intensity, tail length, tail moment) are used to assess DNA damage among organisms exposed to environmental contaminants (e.g., Mincarelli et al., 2019).
- Encephalization index, ratio of brain to body weight among species. Used to compare cognitive abilities.

Scales

Agreement scales for surveys, e.g., **Likert scale** or sliding scale (Sullivan and Artino 2013). For example, after learning about Theranos, students were asked:

How serious is this violation in your opinion (on a 5-point scale)?

Not serious	Slightly serious	Moderately serious	Serious	Very serious
0	0	2	4	19

Although an intuitive measure, how fast an individual can run is challenging to determine because it is difficult to ensure that an individual's performance is at physiological maximum. Measures of performance capacity that involve behavior (motivation) can be particularly challenging, which may lead to the use of a **race quality** scale (eg., binary scale "good" or "bad" Husak et al 2006).

These examples reflect ordinal scales. Many of the nonparametric tests discussed in Chapter 15 are suitable for analysis of scales.

Limitations of ratios

Although the indexes may be easy to communicate, statistically, indexes have many drawbacks. Chief among these is that variation in ratios may be due to change in numerator or denominator. Ratios and any index calculated by combining ratios seem simple enough, but have complicated statistical properties. Over the years, several authors have made critical suggestions for use of ratios

and indexes. Some key references are Packard and Boardman (1988), Jasienski and Oikos (1999), Nee et al (2005), and Karp et al (2012). For example, ratios, computing trait value by body weight, are often used to compare some trait among individuals or species that differ in body size. However, this **normalization** attempt only removes the covariation between size and the trait if there is a 1:1 relationship between size and the trait. More typically, relationship between the trait and body size is allometric, i.e., the slope is not equal to one. Thus, ratio will over-correct for large size and under-correct for small size. The proper solution is to conduct the comparison as part of an analysis of covariance (ANCOVA, see Chapter 17.6).

Example

Which is the safer mode of travel: car or airplane?

The following discussion covers travel safety in the United States of America for a typical year, 2000*.

Note:

*Note that the following discussion excludes the 241 airline passenger deaths associated to the terrorist attacks of September 11, 2001 in the USA; the NTSB also "…exclude(s illegal acts) for the purpose of accident rate computation." It also does not include considerations of 2020–2021 and effects of the COVID-19 pandemic on numbers of flights. The purpose of this discussion is not to convince you about the safety of modes of travel. Moreover, the following analysis is not necessarily the proper way to frame or analyze risk, but, rather, the purpose of this discussion is to highlight the impact of assumptions on estimating risk.

Between 2000 and 2023, there were 779 deaths associated with accidents of major air carriers in the USA. Year 2009 was the last multiple-casualty crash of a major U.S. carrier (Colgan Air Flight 3407); between 2010 and 2021, two fatal accidents, two fatalities were reported.

We've all heard the claim that it's much safer to fly with a major airline than it is to travel by car (e.g., 1 January 2012 article in online edition of San Francisco Chronicle). There are a variety of arguments, but one statistical argument goes as follows. In 2000 in the United States, 638,902,993 persons traveled by major air carrier, whereas there were 190,625,023 licensed drivers. In 2000, 92 persons died in air travel (again, major carriers only), whereas 37,526 persons died in vehicle crashes (includes drivers and passengers). Thus, the risk of dying in air travel is given as the proportion $\frac{92}{638902993}$, or 1.44×10^{-7} (0.000014%), whereas the comparable proportion for death by motor vehicle is $\frac{37526}{190625023}$, or 1.97×10^{-4} (0.0197%).

In other words, we can expect one death (actually 1.4) for every ten million airline passengers, but 20 deaths (actually 19.7) for every one hundred thousand licensed drivers. Thus, flying is a thousand times safer than driving (actual result 1,367 times; divide the rate of motor vehicle-caused deaths for licensed drivers by the rate for airlines). Proportions are hard to compare sometimes, especially when the **per capita** numbers differ (ten million vs. 100,000 in this case).

We can put the numbers onto a **probability tree** and get a sense of what we are looking at.

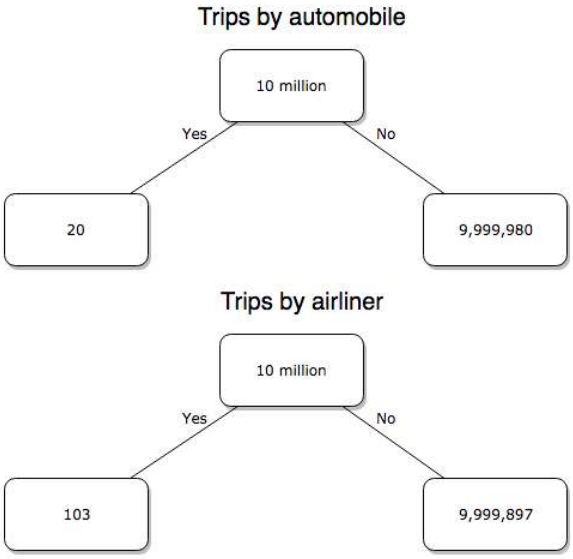


Figure 6.2.1: A **probability tree** to help visualize comparison of deaths ("yes") by car travel and by airline travel in the United States for the year 2000.

Comparing rates and proportions

Without going into the details, we will do so in Chapter 9: Inferences on Categorical Data, comparing two rates is a **chi-square**, χ^2, **contingency table** type of problem. More specifically, however, it is a **binomial** problem (Chapter 3.1, Chapter 6.5); there are two outcomes, death or no death, and we can describe how likely the event is to occur as a proportion. Because the numbers are large, we can use rely on the **normal distribution** for comparing the two proportions. We'll explain this more in the next chapters, but for now it may be enough to present the equation for the comparison of two proportions under the assumption of normality, **proportion z test**.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and the null hypothesis (see Chapter 8) tested as that the two proportions are equal. This may be written as

$H_0 : p_1 - p_2 = 0$

We can assign statistical significance to the differences in events for the two modes of travel under this set of assumptions. `Rcmdr` has a nice menu-driven system for comparing proportions, but for now I will simply list the R commands.

At the R prompt, type each line then submit the command.

```
total = 100000000
prop.test(c(19700,14),c(total,total)))
```

And the R output is:

```
prop.test(c(19700,14),c(total,total))

    2-sample test for equality of proportions with continuity
    correction

data:  c(19700, 14) out of c(total, total)
X-squared = 19658, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
 0.0001940984 0.0001996216
sample estimates:
  prop 1   prop 2
1.97e-04 1.40e-07
```

There's a bit to unpack here. R is consistent; when it reports results of a statistical test, it typically returns the value of the test statistic $\left(\chi^2 = 19658\right)$, the degrees of freedom for the test (df = 1), and the p-value (< 2.2e-16).

Note:

The confidence intervals reported by `prop.test()` were calculated by the **Wilson Score method**, not the **Wald method**. While both are parametric tests and therefore sensitive to departures from normality (see Chapter 13.3), formulation of Wilson score method makes fewer assumptions (involving approximations of the population proportions) and therefore is considered more accurate.

By convention in statistics, if a p-value, where "p" stands for probability, is less than 5%, we would say that our results are statistically significant from the null hypothesis. Looks pretty convincing to me; the difference of 19,700 deaths compared to 14 deaths is clearly different by any criterion and by the results of the statistical test, the p-value is several **orders of magnitude** smaller than 5%.

Note:

Order of magnitude generally refers to differences in multiples of ten, logarithmic: a difference of one order of magnitude is the number multiplied by 10^1, three orders of magnitude is the number multiplied by 10^3, and so on.

Safer to fly. By far, not even close. And similar conclusions would be reached if we compare different years, or averages over many years, or if we used a different way to express the amount of travel (e.g., miles/year) by these modes of transportation.

Are you convinced, really? Is it safer to fly?

Let's try a little statistical reasoning — what **assumptions** did I make to do these calculations? We recognize immediately that many more people travel by car: that there are way more cars being driven then there are airline planes being flown. The question then is, have we properly adjusted for this difference? Here are a few considerations. My source for the numbers is the *NTS 2001* book published by the U.S. Department of Transportation (www.dot.gov). We are conducting a risk analysis, and the first step is to make sure that we are comparing "apples with apples." Here are two alternative solutions that at least compare, "Red Delicious" apples with "Macintosh" apples.

Option 1

There are many, many more licensed drivers than there are licensed commercial airline pilots. The standard comparison offered in the background above compared deaths per licensed car driver, but a different metric for air travel, the rate per passenger. This isn't as bad of a comparison as it may seem — after all, the majority of deaths in car accidents are of the driver themselves. But it isn't that hard to make the direct comparison — just find out how many commercial pilots there are — a direct comparison with licensed car drivers (stated above as 190,625,023). From the FAA we see that in 2009 there were 125,738 persons with commercial certificates. Since there are only 20 major airline carriers in the United States now (a few more were active in 2000, but we'll put this aside), the number of licenses is an overestimate of the actual number we want — how many pilots of commercial airlines — but let's use this number for starters. After all, just because a person has a drivers license doesn't mean they drive or ride in a car.

Number of deaths/yr: Let's use 2000 data, a typical year prior to 9/11 (and excluding the Covid-19 pandemic). Airlines: 92 deaths; motor vehicles (includes passenger cars, trucks, etc., but not motorcycles): 37,526 deaths (drivers = 25,567; passengers = 10,695; 86 others).

Which mode of travel is riskier? I get a rate a rate of 7.3×10^{-4} deaths per commercial pilot, compared to a rate for car drivers of 1.97×10^{-4} deaths.

To summarize what we have so far, I get a result that suggests car travel is almost four times safer:

$$\frac{7.3 \times 10^{-4}}{1.97 \times 10^{-4}} \rightarrow \frac{7.3}{1.97} = 3.7$$

then traveling by commercial airliner. In whole numbers, these results translate to seven deaths for every 10,000 commercial pilots compared to two deaths for every 10,000 licensed car drivers.

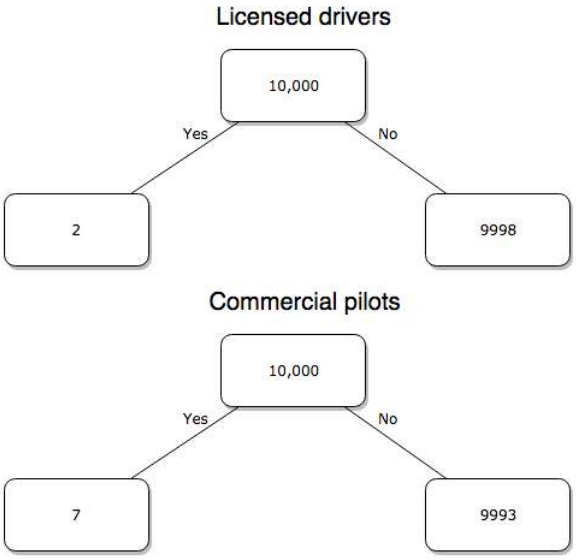


Figure 6.2.2: Comparing totals of deaths adjusted by numbers of licensed drivers and by licensed commercial airline pilots in the United States.

R work follows. Enter and submit each command on a separate line in the script window

```
total = 10000
prop.test(c(2,7),c(total,total))
```

And the R output

```
prop.test(c(2,7),c(total,total))

    2-sample test for equality of proportions with continuity
    correction

data:  c(2, 7) out of c(total, total)
X-squared = 1.7786, df = 1, p-value = 0.1823
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.001187816  0.000187816
sample estimates:
prop 1 prop 2
 2e-04  7e-04
```

What's happened? The p-value (0.1823) is not less than 5%, and so we would conclude under this scenario that there is no difference between the proportions of deaths between the two modes of travel. Let's keep going.

Option 2

There are many, many more cars on the road then there are airplanes flying commercial passengers. The standard comparison offered in the background information above identified death rates per individual driver, but used a different metric for airline

travelers (number of deaths per passenger), which confuses individuals with travelers: what we need is the number of individuals that traveled by airliner, not the total number of passengers (which is many times higher, because of repeat flyers). How can we make a fair comparison for the two modes of travel? Most people never fly, whereas most people drive (or ride in a car) frequently in the United States. To me, risk of travel might be better expressed in terms of a per trip rate. I want to know, what are my chances of dying each time I get into my car versus each time I fly on a commercial jet in the United States?

Number of trips/yr. For airlines, I use the number of departures (in 2000 this was 8,951,773). But for cars, we need to decide how to get a similar number. It's not available directly from the DOT (and would be difficult to get — studies with randomly selected drivers can yield as many as 5 trips per day for licensed drivers). I took the number of licensed drivers and bound the problem — at the low end, let's say that only 2 trips per week (e.g., 50 weeks) are taken by licensed drivers (100 trips); at the upper end, let's take 2 trips per day per week, or 500 trips/year. Thus, at the low end, we have 1.91×10^{10} trips per year; at the upper end, 9.53×10^{10} trips per year.

Which mode of travel is riskier? Using the number of deaths/yr listed above in Option 1, I get a rate of 1.03×10^{-5} deaths per trip for air carriers compared to a rate of 1.97×10^{-6} deaths per trip for cars (lower bound) or 3.9×10^{-7} deaths per trip for cars (upper bound). Here's what the numbers look for in a tree (taking the lower number of trips per year for cars).

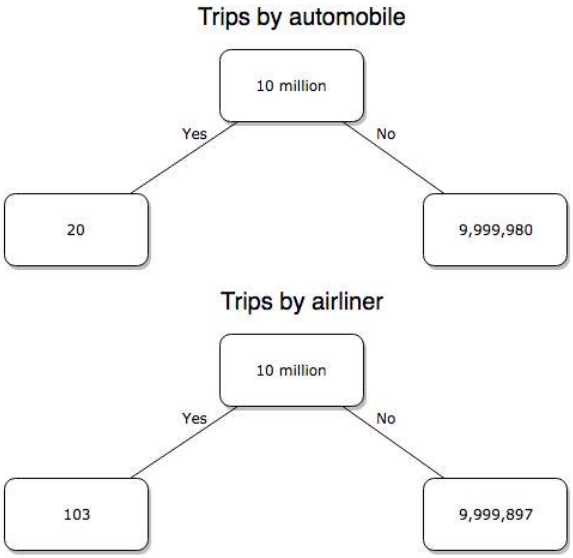


Figure 6.2.3: Comparing totals of deaths adjusted by numbers of car trips and by numbers of airline trips in the United States.

R work follows:

```
total = 10000000
prop.test(c(20,103),c(total,total))
```

And the R output:

```
prop.test(c(20,103),c(total,total))

    2-sample test for equality of proportions with continuity
    correction

data:  c(20, 103) out of c(total, total)
X-squared = 54.667, df = 1, p-value = 1.428e-13
alternative hypothesis: two.sided
95 percent confidence interval:
 -1.057370e-05 -6.026305e-06
```

```
sample estimates:
  prop 1   prop 2
2.00e-06 1.03e-05
```

Now we have another really small p-value $\left(1.428 \times 10^{-13}\right)$, which suggests a statistically significant difference between the modes of travel, but the difference in deaths is switched. I now have a result that suggests car travel is *much* safer then traveling with a commercial airliner! These calculations suggest that you are as much as 26 (upper bounds, five times for lower bounds) times more likely to die from a plane crash then you are behind the wheel. In whole numbers, these results indicate one death for every 100,000 airline flights compared to 1 death for every 500,000 (lower estimate) or 2,500,000 car trips!

Do I have it right and the standard answer is wrong? As Lee Corso says often on ESPN's College GameDay program, "Not so fast, my friend!" (Wikipedia). Mark Twain was right to hold the skeptic's view. Begin by listing the assumptions and by checking the logic of the comparisons (there are still holes in my logic!!). For one, if I am considering my risk of dying by mode of travel, it is far more likely that I will be in a car accident than I will an airline accident, simply because I don't travel by airline that much. When we consider **lifetime risk**, we can see why the assertion that it is "safer to fly than drive" is true — we're far more likely to belong to one of the reference populations involving automobiles (e.g., those who drive frequently, for many years) than we are to be among the frequent flyers reference populations.

Questions

1. Review and provide your own examples for

- index
- rate
- ratio
- proportion

2. Return to my story about travel safety, airlines vs cars: am I using "statistic" or "statistics?"

3. Like travel safety, we are often confronted by risk comparisons like the following: Which animal is more deadly to humans, dogs or sharks? Between the two, which lead to more hospitalizations in the United States? Work through your assumptions and use results from the International Shark Attack file.

- If a person lives in Nebraska, and never visits the ocean, how does a "shark attack" risk analysis apply? Is it a fair comparison to make between dog attacks and shark attacks? Why or why not.

4. Go to cappex.com/colleges and update institutional (gift) aid offered by Chaminade and HPU. Compare to University of Hawaii-Manoa.

6.3: Combinations and permutations

Events

The simplest place to begin with probability is to define an **event** — an event is an outcome — a "Heads" from a toss of coin, or the chance that the next card dealt is an ace in a game of Blackjack. In these simple cases we can enumerate how likely these events are given a number of chances or trials. For example, if the coin is tossed once, then the coin will either land "Heads" or "Tails" (Fig. 6.3.1).

Figure 6.3.1: Heads (left) and Tails (right) of a USA quarter.

We can then ask, what is the likelihood of ten "Heads" in ten tosses of a fair coin?

And this gets us to some basic rules of probability:

- if successive events are independent, then we multiply the probability
- if events are not independent, then the probability adds
- do we sample without replacement, i.e., an object can only be used once?
- do we sample with replacement, i.e., an object gets placed "back in the deck" and can be used again and again?

Combinations and Permutations

Combinations ignore the order of the events; **permutations** are ordered combinations. For replacement, the formula for permutations is simply n^r

In the case of the ten "Heads," in ten successive trials, the probability is $1/2$ "ten times" or $\left(\frac{1}{2}\right)^{10} = 0.0009766$ (in R, just type `0.5^10` or `(1/2)^10` at the R prompt).

Examples of permutations in biology include:

- Given the four nucleotide bases of DNA, adenine (A), cytosine (C), guanine (G), and thymine (T), how many codons are possible?

  $$4^3 = 64$$

  where the three (3) refers to the codon, a three-nucleotide genetic code. Codons are trinucleotide sequences of DNA or RNA that correspond to a specific amino acid.
- How often do we expect to find by chance the heptamer (i.e., seven-base) consensus TATA box sequence, TATAWAW (W can be either Adenine or Thymine, per Standard IUB/IUPAC nucleic acid code)?

  $$4^7 = 16384$$

  Thus, we would expect at random to find TATA box sequences every 16,384 bases along the genome. For the human genome of 3.3 billion bases, then we would expect at random more than 200,000 TATA box consensus sequences.

Another way to put this is to ask how many combinations of ten tosses gets us ten "Heads" then weigh this against all possible outcomes of tossing a coin ten times — how many times do we get zero Heads? Two "Heads"? And so on. This introduces the combinatorial.

$$C(n,r) = \frac{n!}{(n-r)!r!}$$

where n is the number of choices (the list of events) and r is how many times you choose or trials. The exclamation point is called the **factorial**. The factorial instructs you to multiply the number over and over again, "down to one." For example, if the number is 10, then

$$10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

In R, just type `factorial(10)` at the R prompt.

```
factorial(10)
[1] 3628800
```

An example

Consider an Extrasensory Perception, ESP, experiment. We place photocopies of two kinds of cards, "Queen" and "King," into ten envelopes, one card per envelope (Fig. 6.3.2). Thus, an envelope either contains a "Queen" or a "King."

Figure 6.3.2: Playing cards with images commemorating 150th anniversary of Charles Darwin's *Origin of Species*. (Design John R. C. White, Master of the Worshipful Company of Makers of Playing Cards 2008 to 2009.)

We shuffle the envelopes containing the cards and volunteers guess the contents of the envelopes, one at a time. We score the correct choices after all envelopes had been guessed. (It's really hard to set these kinds of experiments up to rule out subtle cues and other kinds of experimental error! Diaconis 1978, Pigliucci 2010, pp. 77–83.) By chance alone we would expect 50% correct (e.g., one way to game the system would have been for the volunteer simply to guess "Queen" every time; since we had placed five "Queen" cards the volunteer would end up being right 50%).

What are the possible combinations?

Let's start with one correct; the volunteer could be right by guessing the first one correctly, then getting the next nine wrong, or equivalently, the first choice was wrong, but the second was correct followed by eight incorrect choices, and so on.

We need math!

Let $n = 10$ and $r = 1$.

Using the formula above, we have

$$C(10, 1) = \frac{10!}{(10-1)!1!}$$

or ten ways to get one out of ten correct. This is the number of combinations without replacement.

In R, we can use the function `choose()`, included in the base install of R. At the R prompt type

```
choose(10,1)
```

and R returns

```
[1] 10
```

For permutations, use `choose(n,k)*factorial(k)` .

To get permutations for larger sets, you'll need to write a function or take advantage of functions written by others and available in packages for R. See `gtools` package, which contains programming tools for R. There are several packages that can be used to expand capabilities of working with permutations and combinations. For example, install a package and library called `combinat` and then run the `combn()` function, together with the `dim()` function, which will return

```
dim(combn(10,1))[2]
[1] 10
```

R explained

In the `combn()` the "10" is the total number of envelopes and the "1" is how many correct guesses. We also used the `dim()` function to give us the size of the result. The `dim()` function returns two numbers, the number of rows and the number of columns — `combn()` returns a matrix and so `dim()` saves you the trouble of counting the outcomes — the "[2]" tells R to return the total number of columns in the matrix created by `combn()` . Here's the output from `combn()` , but without the `dim()` function

```
combn(10,1)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    2    3    4    5    6    7    8    9    10
```

in order, we see that there is one case where the success came the first time [1,1], another case where success came on the second try `[1,2]` and so on.

We can write a simple function to return the combinations

```
for (many in seq(0, 10, by = 1)){
    print(choose(10, many))
}
```

an inelegant function, but it works well enough, returning

```
[1] 1
[1] 10
[1] 45
[1] 120
[1] 210
[1] 252
[1] 210
[1] 120
[1] 45
[1] 10
[1] 1
```

Note that there's only one way to get zero correct, only one way to get all ten correct.

For completion, what would be the permutations? Use the `permn()` function (same `combinat` library) along with `length()` function to get the number of permutations

```
length(permn(10))
[1] 3628800
```

Continue the example

Back to our ESP problem. If we continue with the formula, how many ways to get two correct? Three correct? And so on, we can show the results in a bar graph (Fig. 6.3.3).

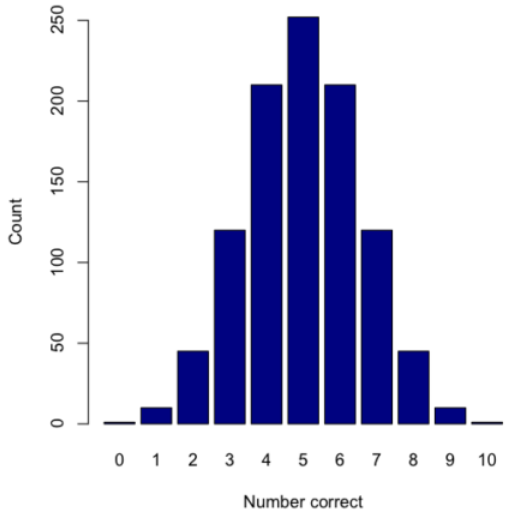


Figure 6.3.3: Bar chart of the combinations of correct guesses out of 10 attempts (graph was presented in Chapter 4.1).

The graph in Figure 6.3.3 was made in R:

```
Combos <- seq(0,10, by=1)
HowMany <- c(1,10,45,120,210,252,210,120,45,10,1)
barplot(HowMany, names.arg = Combos, xlab = "Number correct", ylab = "Count",col = "d
```

If you recall, we had two students score eight out of ten. Is this evidence of "ESP?" It certainly seems pretty good, but how rare is this to score 8 out of 10? The total number of outcomes was

$$1+10+45+\cdots+45+10+1=1024 \text{ ways}$$

Eight out of ten could be achieved in 45 different ways, so

$$\frac{45}{1024}=0.043945$$

So, eight out of ten is pretty unlikely! Is it evidence for ESP powers in our Biostatistics class? In fact, at the traditional Type I error rate of 5% used in biology and other sciences to evaluate inferences from statistical tests (See Chapter 8), we would say that this observation was statistically significant. Given that this would be an extraordinary claim, this should give you an important clue that statistical significance is not the same thing as evidence for the claim of ESP. In other words, Is the result biologically significant? Probably not, but I'll keep my eyes on you, just in case.

Conclusions

Combinations or permutations?

Combinations refers to groups of n things taken k times without repetition. Note that order does not matter, just the combination of things. Permutations, on the other hand, specifically relate the number of ways a particular arrangement can show up.

Questions

1. Calculate the combinations, from zero correct to ten correct, from our ESP experiment, i.e., confirm the numbers reported in Figure 6.3.3.
2. Consider our ESP tests based on guessing cards. Let's say that one subject repeatedly reports correct guesses at a rate greater than expected by chance. Why or why not should we view this as evidence the person may have extrasensory perception?
3. Consider a common DNA triplet repeat, the three letter *CAG*.
    1. How many permutations are there for this triplet (word)?
    2. How many combinations are there for this triplet (word)?
4. We call them combination locks, but given the definition of combination, is that the correct use of the term? Explain (and for those of you who insist on searching for "the answer," cite your sources).

6.4: Types of probability

Introduction

By **probability**, we mean a number that quantifies the uncertainty associated with a particular event or outcome. If an outcome is certain to occur, the probability is equal to 1; if an outcome cannot happen, the probability is equal to zero. Most events we talk about in biology have probability somewhere in between.

A basic understanding, or at least an appreciation for probability is important to your education in biostatistics. Simply put, there is no certainty in biology beyond simple expectations like Benjamin Franklin's famous quip about "death and taxes…"

Discrete probability

You are probably already familiar with **discrete probability**. For example, what is the probability that a single toss of a fair coin will result in a "heads?" The outcomes of the coin toss are discrete, categorical, either "heads" or "tails."

> **Note:**
>
> Obviously, this statement assumes a few things — coin tossed not in a vacuum; although possible, we ignore the possibility of the coin landing on edge.

And for ten tosses of the coin? And more cogently, what is the probability that you will toss a coin ten times and get all ten "heads?" While different in tone, these are discrete outcomes. An important concept is independence. Are the multiple events independent? In other words, does the toss of coin on the first attempt affect the toss of the coin on the second attempt and so on up to the tenth toss? At least in principle, the repeated tosses are independent, so to find the probability you just multiply each event's probability to get the total. In contrast, if one or more events are not independent but somehow influence the behavior of the next event, then you add the probabilities for each dependent event. We can do better than simply multiply or add events one at a time; depending on the number of discrete outcomes, it is very likely that someone has already calculated all possible outcomes and come up with an equation. In the case of the tossing of two coins, this is a binomial equation problem and repeat tosses can be modeled by use of the Bernoulli distribution.

Now, try this on for size. What is the probability that the next child born at Kapiolani Medical Center in Honolulu will be assigned female?

We just described a discrete random variable, which can only take on discrete or "countable" numbers. This distribution of values is the **probability mass function**. The probability of one fatal airline accident in a year exactly on 20.1 is practically zero (the area under a point along the curve is zero), so we can get the probability of a range of values around the point as our answer.

Continuous probability

Many events in biology are of degree, not kind. It is kind of awkward to think about it, but for a sample of adult house mice drawn from a population, what is the probability of obtaining a mouse that is 20.0000 grams (g) in weight? Each possible value of body mass for a mouse is considered an event, just like in our example of tossing a coin. But clearly, we don't expect to get a lot of mice that are exactly 20.0000 g in weight. For variables like body mass, the type of data we collect is continuous, and the probability values need to be rethought along a continuum of possible values and, in turn, how likely each value is for a mouse. Although it is theoretically possible that a mouse could weigh ten pounds, we know by experience that this is impossible. Adult mice weigh between 15 and 50 g or thereabouts.

We just described a continuous random variable, which can take on any value between a specific interval of values. This distribution of values is the **probability density function**. The probability of a mouse's weight falling exactly on 20.0 is practically zero (the area under a point along the curve is zero), so we can get the probability of a range of values around the point as our answer.

In statistical inference, following our measurements of the variables from our sample drawn from a population, we make conclusions with the following kind of caveat: "the mean body mass for this strain of mouse is 20 g." That is our best estimate of the mean (middle) for the population of mice, more specifically, for the body mass of the mice. Here, the variable body mass is more formally termed a **random variable**. This implies that there is in fact a true population mean body mass for the mice and that any deviations from that mean are due to chance. In statistics we don't settle for a single point estimate of the population mean. You will find that most reporting of estimates of random variables is accompanied by a statement like "the mean was 20 g with a

95% probability that the true population mean is between 18.9 and 21 g." This is called the 95% **confidence interval** for the mean and it takes into account how good of an estimate our sample is likely to be relative to the true population value. Not only are we saying that we think the population mean is 20 g, but we're willing to say that we are 95% certain that the true value must be between a lower limit (18.9 g) and and upper limit (21 g). In order to make this kind of statement, we have to assume a distribution that will describe the probability of mouse weights. For many reasons we usually assume a normal distribution. Once we make this assumption we can calculate how probable a particular weight is for a mouse.

We introduced how to calculate Confidence Intervals in Chapter 3.4 and will extend this in Chapter 7.6 and Chapter 8.6.

Types of probability

To begin refining our concept of probability, it is sometimes useful to distinguish among kinds of probabilities:

- between **theoretical** and **empirical**;
- between **subjective** and **objective**.

In most cases, including your statistics book, we would begin our discussion of probability by talking of some probabilities for events we're familiar with.

1. The theoretical probability of heads appearing the next time you flip a fair coin is 1/2 or 50%. As long as we're talking about a fair coin, the probability of a heads appearing each time you flip the coin remains 50%. We can check this by conducting and experiment: out of 10 tosses, how many heads appear? The answer would be an empirical probability, and we understand the chance in an objective manner (no interpretation needed).
2. The theoretical probability that a "5" will appear on the face of a fair die after a toss is 1/6 or 16.667%. Again, as long as we're talking about a fair, standard 6-sided die, the probability of a "5" appearing each time you roll the dice remains 16.667%.
3. The probability that at birth, a human baby's sex will be male is about 1/2 or 50%. This is an empirical probability based on millions of observations. Changes in technology and ethical standards notwithstanding, the probability will remain the same.
4. The probability of the birth of a Downs syndrome baby is 1/800, but increases with age of the mother until by age 45, the chance is 1/12. Again, these are empirical and objective.
5. The probability of winning the Publisher's Clearing House Sweepstakes is about 1 in 100 million. This probability is theoretical, it is also objective; however, by adding lots of twists to the game, by having multiple opportunities and by giving the appearance that a person must purchase a magazine, some players perceive their chances as increasing or decreasing by their efforts (=subjective).

R and distributions

R Commander: Distributions menus give four options

- Quantiles
- Probabilities
- Plot
- Sampling

Questions

1. Define and distinguish, with examples

1. discrete and continuous probability
2. theoretical and empirical probability
3. subjective and objective probability

6.5: Discrete probability distributions

Binomial distribution

Discrete refers to particular outcomes. Discrete data types include all of the categorical types we have discussed, including binary, ordinal, and nominal.

The binomial probability distribution is a discrete distribution for the number of successes, k, in a sequence of n independent trials, where the outcome of each trial can take on only one of two possible outcomes. For cases of 0 or 1, yes or no, "heads" or "tails," male or female, we talk about the binomial distribution, because the outcomes are discrete and there can be only two possible (**binary**) outcomes.

> **Note:**
>
> Fair coins have two sides; tossing a coin we expect "heads" or "tails," but rarely, some coin types (e.g., USA nickels) may land and come to rest on edge or side. We still consider the coin toss having binary outcomes, by definition, even though a coin may land on edge about one toss in six thousand (Murray and Teare 1993) because the exception is extremely rare. h/t Dr. Jerry Coyne.

The mathematical function of the binomial is written as

$$Pr[X\ successes] = \left(\frac{n}{X}\right) p^X (1-p)^{n-X}$$

where the **binomial coefficient** is given by

$$\left(\frac{n}{X}\right) = \frac{n!}{X!(n-X)!}$$

and X refers to the number of ways to choose "success" from n observations.

Consider an example.

We have to define what we mean by success. For coin toss, this might be the number of heads.

The mean for the binomial this is given simply as

$$\mu_X = np$$

where X is "Heads" (the category of successes for our example), and p corresponds to the probability the selected event occurs, in this case, "Heads."

The variance of the binomial distribution is given by

$$\sigma_X^2 = np(1-p)$$

Here's a density plot of two trials with success 2% with $n(x)$ equal to 20 (Fig. 6.5.1).

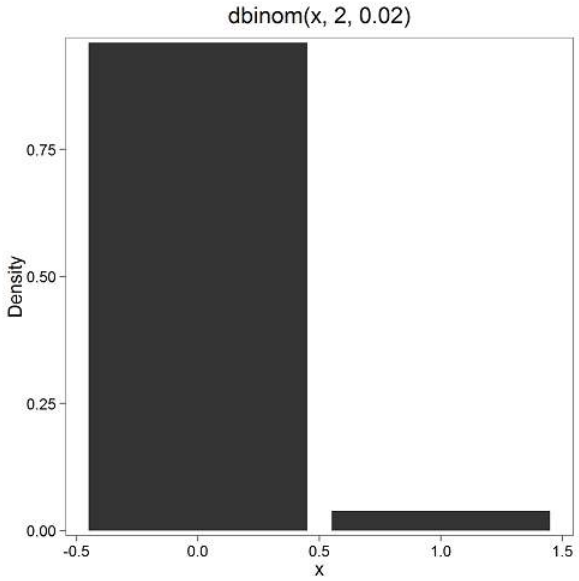


Figure 6.5.1: Plot generated with KMggplot2 Rcmdr plugin.

Here's the R code.

Create the trials, 1 through 20, then create an object to hold the number of trials:

```
nSize=1:20
Size <- length(nSize); Size
```

R returns:

```
[1] 20
```

Assign the probability value to an object:

```
prob <- 0.02
```

Next, calculate the mean, mu, and the variance, var, for the binomial with prob = 0.02 and the number of trials as Size = 20:

```
mu <- Size*prob
var <- Size*prob*(1-prob)
```

Print the mean and variance; let's assign them to an object then print the object:

```
stats <- c(mu, var); stats
```

And R returns:

```
[1] 0.400 0.392
```

And here's a real-world example. Twinning in humans is rare. In Hawaii in the 1990s the rate of twin births (monozygotic and dizygotic) was about 20 for every 1000 births or 2%. "Success" here then is twin births.

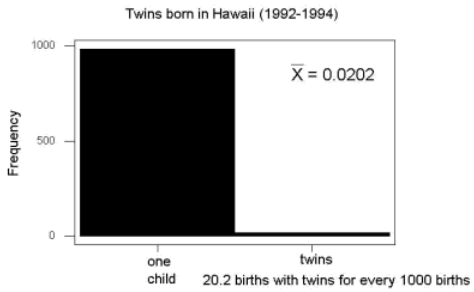


Figure 6.5.2: Example of binomial-like distribution: reported twins born in Hawaii.

Interestingly, rates of twins have since increased in Hawaii (31 out of 1000 births) and in the United States overall (33 out of 1000 births) (Table 2, NCHS Data Brief No. 80, 2012). Data were for year 2009.

Out of 10 births, what is the probability of two twin births in Hawaii?

$$Pr[2\ twin\ births] = \left(\frac{10}{2}\right) 0.031^2 (1-0.031)^{10-2}$$

You can solve this with your calculator (yikes!), or take advantage of online calculators (GraphPad QuickCalcs), or use R and Rcmdr.

In R, simply type at the prompt

```
dbinom(2,10,0.031)
[1] 0.03361446
```

Try in R Commander.

Rcmdr → Distributions → Discrete distributions → Binomial distribution → Binomial probabilities …

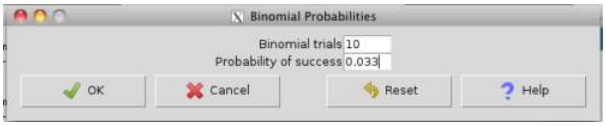


Figure 6.5.3: Rcmdr menu to get binomial probability.

Note I used $p = 0.033$, the rate for the entire USA. Here's the output.

```
> .Table <- data.frame(Pr=dbinom(0:10, size=10, prob=0.033))
> rownames(.Table) <- 0:10
> .Table
Pr
0  7.149320e-01
1  2.439789e-01
2  3.746728e-02 ← Answer, 0.0375 or 3.75%
3  3.409639e-03
4  2.036263e-04
5  8.338782e-06
6  2.371422e-07
7  4.624430e-09
8  5.918028e-11
9  4.487991e-13
10 1.531579e-15
```

And here is the output for our example from Hawaii ($p = 0.031$).

```
> .Table <- data.frame(Pr=dbinom(0:10, size=10, prob=0.031))
> rownames(.Table) <- 0:10
> .Table
Pr
0  7.298570e-01
1  2.334940e-01
2  3.361446e-02 ← Answer, 0.0336 or 3.36%
3  2.867694e-03
4  1.605494e-04
5  6.163507e-06
6  1.643178e-07
7  3.003893e-09
8  3.603741e-11
9  2.561999e-13
10 8.196283e-16
```

We use the binomial distribution as the foundation for the **binomial test**, i.e., the test of an observed proportion against an expected population level proportion in a Bernoulli trial.

Hypergeometric distribution

The binomial distribution is used for cases of **sampling with replacement** from a population. When **sampling without replacement** is done, the **hypergeometric distribution** is used. It is the number of successes, k, in a sequence of n independent trials drawn from a fixed population. This sampling scheme means that each draw is no longer independent — with each draw you decrease the remaining number of observations and thus change the proportion.

The mathematical function of the hypergeometric is written as

$$Pr[X = k] = \frac{\left(\frac{K}{k}\right)\left(\frac{N-K}{n-k}\right)}{\left(\frac{N}{n}\right)}$$

where N is the population size, K is the number of successes in that population, and n and k are defined as above. Let's look apply this to the twinning problem.

In 2009, 2200 women gave birth in Hawaii County, Hawaii. Out of 10 births, what is the probability of 2 twin births in Hawaii?

Assuming "risk" of twinning is the same rate as in rest of USA, then we have expected 72 successes in this population (0.033×2200).

Here's the graph (Fig. 6.5.4),

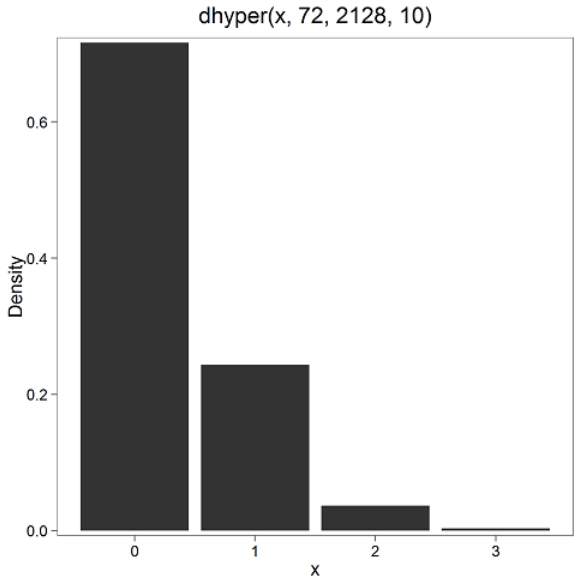


Figure 6.5.4: Plot of hypergeometric distribution of twinning in Hawaii.

where the X axis values shows the number of events with successes (twin births). Taking the bin 2 (we wanted to know about the probability of 2 out of ten), we can draw a line back to the Y-axis to get our probability — looks like roughly 5%. Plot drawn with KMggplot2.

To get the actual probability,

Rcmdr → Distributions → Discrete distributions → Hypergeometric distribution → Hypergeometric probabilities …

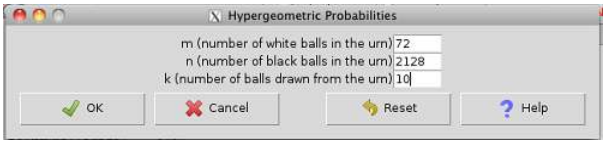


Figure 6.5.5: Rcmdr menu to get hypergeometric probability.

where m is the number of successes, n is the number of "failures," and k is the number of trials.

```
> .Table
Pr
0 0.716453457
1 0.243438645
2 0.036688041 ← Answer, 0.0367 or 3.67%
3 0.003228871
```

The reference to white and black balls and urns is a device described by Bernoulli himself and has been used by others ever since to discuss probability problems (called the urn problem), and so I apply it here to be consistent. The urn contains a number of white (x) and a number of black (y) balls mixed together. One ball is drawn randomly from the urn — what color is it? The ball is then is either returned into the urn (replacement) or it is left out (without replacement) as in the hypergeometric problem, and the selection process is repeated.

Besides applications in gambling and balls-in-urns problems, this distribution is the basis for many tests of gene enrichment from microarray analyses. The hypergeometric forms the basis of the **Fisher Exact test** (see Chapter 9.5).

Discrete uniform distribution

For discrete cases of "1," "2," "3," "4," "5," or "6," on the single toss of a fair die, we can talk about the discrete **uniform distribution** because all possible outcomes are equally likely. If you are branded as a "card-counter" in Las Vegas, all you've done is reached an understanding of the uniform distribution of card suits!

One biological example would be the fate of a random primary oocyte in the human (mammal) female — three out of four will become polar bodies, eventually reabsorbed, whereas one in four will develop into a secondary oocyte (egg); the uniform distribution has to do with the counts of the products — each of the four primary oocytes has the same (apparently) chance (25%) of becoming the egg.

The uniform distribution exists also for continuous data types.

Poisson distribution

An extension from the binomial case is that, rather than following success or failure, you may have the following scenario. Consider a wind-dispersed seed released from a plant. If we mark up the area around the plant in grids, we could then count the number of seeds within each grid. Most grids will have no seeds, some grids will have one seed, a few grids may have two seeds, etc. Multiple seeds in grids is a rare event. The graph might look like

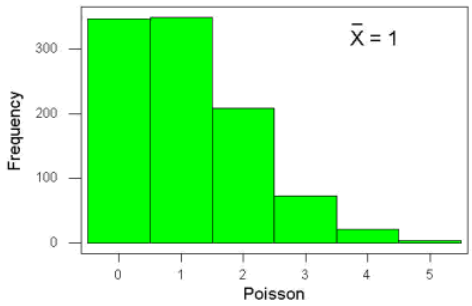


Figure 6.5.6: Example, Poisson-like graph: the number of wind-dispersed seeds within each grid.

The Poisson has interesting properties, one being that the expected mean is equal to the variance. An equation is

$$Pr[X] = \frac{\mu^X e^{-\mu}}{X!}$$

where μ is the mean (or we could substitute with variance!), e is the natural logarithm, and X is number of successes you are interested in. For example, if $\mu = 1$, what is the probability of observing a grid with five seeds? Simple enough to do this by hand, but let's use Rcmdr instead. Here's the graph (Fig. 6.5.7) from Rcmdr (KMggplot2 plugin)

Figure 6.5.7: ggplot2 plot of Poisson distribution, $\mu = 1$.

and for the actual probability we have from R

```
dpois(5, lambda = 1)
[1] 0.003065662
```

Rcmdr → Distributions → Discrete distributions → Poisson distribution → Poisson probabilities … (Fig. 6.5.8)

The only thing to enter is the mean (some call μ lambda with symbol λ).

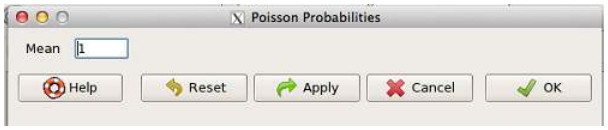


Figure 6.5.8: Rcmdr menu for Poisson probability.

Here's the output from R. For intervals 0, 1, 2, 3, …, 6 (Rcmdr just enters this range for you)!

```
> .Table <- data.fram(Pr=dpois(0:6, lambda = 1))
> rownames(.Table) <- 0:6
> .Table
Pr
0 0.3678794412
```

```
1 0.3678794412
2 0.1839397206
3 0.0613132402
4 0.0153283100
5 0.0030656620 ← Answer, 0.0307 or 3.07%
6 0.0005109437
```

Next — Continuous distributions

And finally, for ratio (continuous) scale data, which can take on any value, we can express the chance that probability of a given point as a continuous function, with the normal distribution being one of the most important examples (there are others, like the F-distribution). Many statistical procedures assume that the data we use can be viewed as having come from a "normally distributed population." See Chapter 6.6.

Questions

1. For each of the following scenarios, identify the most likely distribution that may be assumed:

- Litter size of 100 toy poodle females. A toy poodle is a purebred dog breed: range of litter size is 1 – 4 pups (Borge et al 2011)
- Mean litter size and total number of litters born per season of the year for litters registered within The Norwegian Kennel Club in 2006 and 2007: means by season were Fall 5, Winter 5, Spring 5, Summer 5 (Borge et al 2011)
- C-reactive protein (CRP) blood levels may increase when a person has any number of diseases that cause inflammation. Although CRP is reported as mg/dL, Doctors evaluate a patient's CRP status as all measures below 1.0 are normal, all measures above 1 are above 1.0.

2. Quarterback sacks by game for the NFL team Seahawks, years 2011 through 2022, are summarized below (data extracted from https://www.pro-football-reference.com/).

Sacks	How many games?
0	25
1	46
2	49
3	39
4	25
5	14
6	8
7	2
8	1
9	0

a) Assuming a Poisson distribution, what are the mean (lambda) and variance?

b) The table covers a total of 112 games. How many sacks (events) were observed?

c) What is the probability of the Seahawks getting zero sacks in a game (in 2022, a season was 17 games; prior years a season was 16 games)?

6.6: Continuous distributions

Law of Large Numbers and Central Limit Theorem

Imagine we've collected (sampled) data from a population and now want to summarize the data sample. How do we proceed? A good starting point is to plot the data in a histogram and note the shape of the sample distribution. Not to get too far ahead of ourselves here, but much of the classical inferential statistics demands that we are able to assume that the sampled values come from a certain kind of distribution called the normal, or Gaussian distribution.

Consider a random sample drawn from a normally distributed population of the following series of graphs, Figures 6.6.1– 4:

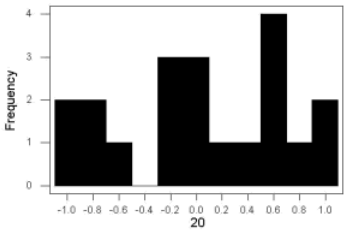


Figure 6.6.1: Sample size = 20, drawn from population with known $\mu = 0$ and $\sigma = 1$.

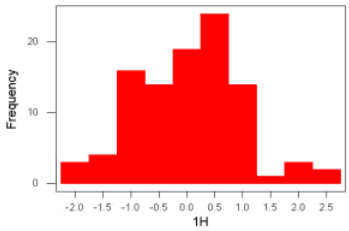


Figure 6.6.2: Sample size = 100, also drawn from population with known $\mu = 0$ and $\sigma = 1$.

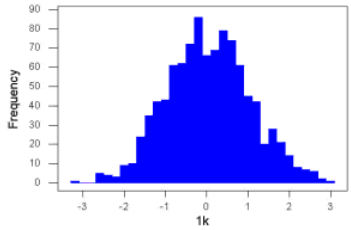


Figure 6.6.3: Sample size = 1000, once again drawn from population with known $\mu = 0$ and $\sigma = 1$.

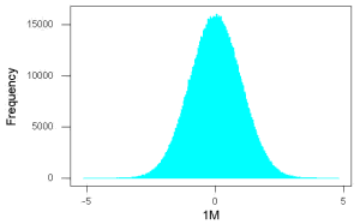


Figure 6.6.4: And lastly, sample size = 1 million also drawn from population with known $\mu = 0$ and $\sigma = 1$.

These graphs illustrate a fundamental point in statistics: for many kinds of measurements in biology, the more data you sample, the more likely the data will approach a normal distribution. This series of simulations was a quick and dirty "proof" of the **Central Limit Theorem**, which is one of the two fundamental theorems of probability, the other being that of **Law of Large Numbers** (i.e., large-sample statistics). Basically the CLT says that for a large number of random samples, the sample mean will approach the population mean, μ, and the sample variance will approach the population variance σ^2; the distribution of the large sample will converge on the normal distribution.

As the sample size gets bigger and bigger, the resulting sample means and standard deviations get closer and closer to the true value (remember — I TOLD the program to grab numbers from the Z distribution with a mean of zero and standard deviation of zero), obeying the Law of Large Numbers.

Simulation

I used the computer to generate sample data from a population. This process is called a **simulation**. R can make new data sets by sampling from known populations with specified distribution properties that we determine in advance — a very powerful tool — a technique used for many kinds of statistics (e.g., Monte Carlo methods, bootstrapping, etc., see Chapter 19).

Note:

How I got the data. All of these data are from a simulation where I asked ~~told~~ R, "grab random numbers from an infinitely large population, with mean = 0 and standard deviation = 1."

1. The first graph is for a sample of 20 points;
2. the second for 100;
3. the third for 1,000;
4. and lastly, 1 million points.

To generate a sample from a normal population, in `Rcmdr` call the menu by selecting:

Rcmdr: **Distributions → Continuous distributions → Normal distribution → Sample from normal distribution…**

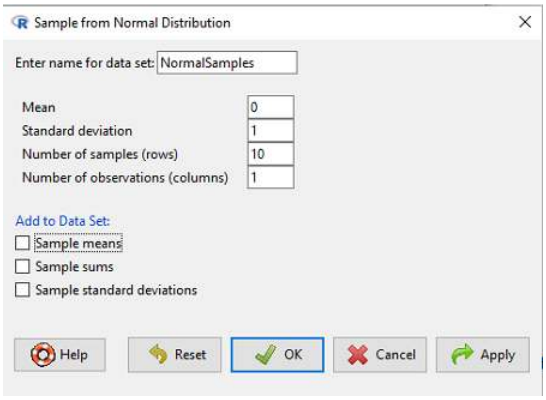


Figure 6.6.5: Screenshot of the Rcmdr menu to sample from a normal distribution.

The menu pops up. I entered Mean $(\mu) = 0$ and Standard deviation $(\sigma) = 1$, number of samples = 10, and unchecked all boxes under the "add to data set." I left the object name as "NormalSamples" but you can, of course, change it as needed. R code derived from these requests were

```
normalityTest(~obs, test="shapiro.test", data=NormalSamples)
NormalSamples <- as.data.frame(matrix(rnorm(10*1, mean=0, sd=1), ncol=1))
rownames(NormalSamples) <- paste("sample", 1:10, sep="")
colnames(NormalSamples) <- "obs"
```

This results in a new `data.frame` called `NormalSamples` with a single variable called `obs` .

Note:

About **pseudorandom number generators**, PRNG. An algorithm is used for creating a sequence of numbers that are like random numbers. We say "like" or "pseudo" random numbers because the algorithm requires a starting number called the **seed**, rather than a truly random process, i.e., a source of entropy outside of the computer. The default PRNG algorithm in base R is Mersenne Twister (Wikipedia), though there are many others included in base R (bring up the help menu by typing `?RNGkind` at the prompt), as well as other packages, like `random` , which can be used to generate **truly random numbers** (source of entropy is "atmospheric noise," per citation in the package, see also random.org).

`rnorm()` was the function used to sample from a normal distribution. If you run the function over and over again (e.g., use a `for loop`), each time you will get different samples. For example, results from three runs

```
for (i in 1:3){
print(rnorm(5))
}
[1] -0.4221672 -1.4317800 -1.8310352 0.4181184 -1.1596058
[1] -0.2034944 1.1809083 1.5925296 -2.0763677 1.6982357
[1] -1.0967218 -0.3205041 -1.7513838 -0.3335311 -1.8808454
```

However, if you set the seed to the same number before calling `rnorm`, you'll get the same sampled numbers.

```
for (i in 1:3){
set.seed(1)
print(rnorm(5))
}
[1] -0.6264538 0.1836433 -0.8356286 1.5952808 0.3295078
[1] -0.6264538 0.1836433 -0.8356286 1.5952808 0.3295078
[1] -0.6264538 0.1836433 -0.8356286 1.5952808 0.3295078
```

The seed number can be any number; it does not have to equal 1.

After creating a sample of numbers drawn from the normal distribution, make a histogram, **Rcmdr**: **Graphs → Histogram…** (see Chapter 4.2).

The normal distribution

More on the importance of normal curves in a moment. Sometimes people call these "bell-shaped" curves. You may also see such distributions referred to as Gaussian Distributions, but the normal curve is but one of many Gaussian-type distributions. Moreover, not all "bell-shaped" curves are NORMAL. For a distribution to be "normally distributed" it must follow a specific formula.

$$Y_i = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(X_i-\mu)^2}{2\sigma^2}}$$

This formula has a lot of parts, but once we break down the parts, we'll see that there are just two things to know. First, let's identify the parts of the equation:

Y_i is the height of the curve (normal density)
π (pi) is a constant = 3.14159… (R, use `pi`)
μ is the population mean
σ^2 is the population variance
σ is the square-root of the variance or the population standard deviation
e is the natural logarithm (R, use `exp()`)
X_i is the individual's value

Why the Normal distribution is so important in classical statistics

With these distinctions out of the way, the first important message about the normal curve is that it permits us to say how likely (i.e., how probable) a particular value is if the observation comes from a population with mean μ and standard deviation σ, and the population from which the sample was drawn came from a normal distribution.

The second message: all we need to know to recreate the normal distribution for a set of data is the mean and the variance (or the standard deviation) for the population!! With just these two parameters, we can then determine the expected proportion of observations expected for each value of X. Note — we generally do not know these two because they are population parameters: we must estimate them from samples, using our sample statistics, and that's where the first big assumption in conducting statistical analyses comes into play!!

Here is an example for calculating the normal distribution when knowing the mean and variance: $\mu = 5$, $\sigma^2 = 10$; thus, the standard deviation is $\sigma = 3.16$.

The formula becomes

$$Y_i = \frac{0.398947}{3.16} \cdot e^{\frac{-(X_i-5)^2}{2 \cdot 10}}$$

Now, plug in different values of X (for example, what's the probability that a value of X could be 0, 1, 2, …, 10 if we really do have a normal curve with mean = 5, and variance = 10?)

The normal equation returns the proportion of observations in a normal population for each X value:

When $i = 5$, $Y_5 = 0.12616$. This is the proportion of all data points that have an $X = 5$ value. When $i = 1$, $Y_1 = 0.019995$. This is the proportion of all data points that have an $X = 1$ value.

We can keep going, getting the proportion expected for each value of X, then make the plot.

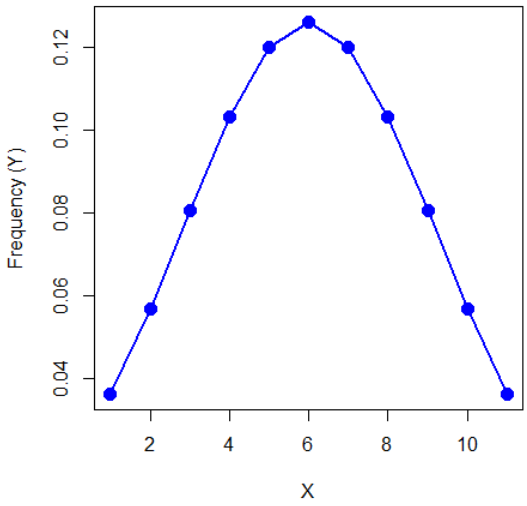


Figure 6.6.6: Frequency expected for a few points ($X = 0$ through $X = 10$) drawn from a normal distribution, calculated using the formula and example values.

Here's the R code for the plot

```
X = seq(0,10, by=1)
Y = (0.398947/3.16)*exp((-1*(X-5)^2)/20)
plot(Y~X, ylab="Frequency (Y)", cex=1.5, pch=19,col="blue")
lines(X,Y, col="blue", lwd=2)
```

Next up is more about the normal distribution, Chapter 6.7.

Questions

1. For a mean of 0 and standard deviation of 1, apply the equation for the normal curve for X = (-4, -3.5, -3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4). Plot your results.
2. Sample from a normal distribution with different sample size, means, and standard deviations. Each time, make a histogram and compare the shape of the histograms.

6.7: Normal distribution and the normal deviate

Introduction

In Chapter 3.3 we introduced the normal distribution and the **Z score**, aka **normal deviate**, as part of a discussion about how some knowledge about characteristics of the dispersion of our data sampled from a population could be used to calculate how many samples we need (the empirical rule). We introduced Chebyshev's inequality as a general approach to this problem, where little is known about the distribution of the population, and contrasted it with the Z score, for cases where the distribution is known to be Gaussian or the normal distribution. The normal distribution is one of the most important distributions in classical statistics. All normal distributions are bell-shaped and symmetric about the mean. To describe a normal distribution only two parameters are needed: the population mean, μ, and the population standard deviation, σ. The normal distribution with mean equal to zero and standard deviation equal to one is called the **standard normal**, or **Z distribution**. With use of the Z score, any normal distribution can be quickly converted to the standard normal distribution.

Proportions of a Normal Distribution

This concept will become increasingly important for the many statistical tests we will learn over the next few weeks. What is the proportion of the populations that is greater than some specific value? Below, again, I have generated a large data set, now with population mean $\mu = 5$ and $\sigma = 2$. The red line corresponds to the equation of the normal curve using our values of $\mu = 5$ and $\sigma = 2$.

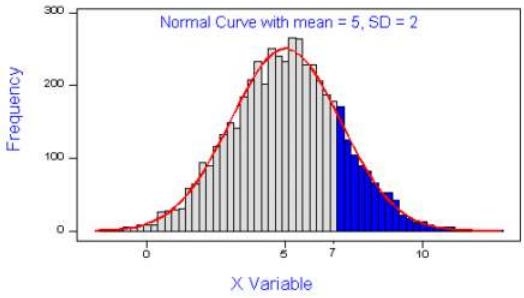


Figure 6.7.1: Frequency of observations expected to be greater than 7, from a large population with $\mu = 5$ and $\sigma = 2$.

Note that this is a crucial step! We assume that our sample distribution is really a sample from a population density (= "area under the curve") function (= "an equation") for a normal random (= "population") variable.

Once I (you) make this assumption, then we have powerful and easy to use tools at our command to answer questions like:

Question: What proportion of the population is greater than 7? (colored in blue).

This gets to the heart of the often-asked question, How many samples should I measure? If we know something about the mean and the variability, then we can predict how many samples will be of a particular kind. Let's solve the problem.

The Z score

We could use the formula for the normal curve (and a lot of repetitions), but fortunately, some folks have provided tables that short-cut this procedure. R and other programs also can find these numbers because the formulas are "built in" to the base packages. First, let's introduce a simple formula that lets us standardize our population numbers so that we can use established tables of probabilities for the normal distribution.

Below, we will see how to use `Rcmdr` for these kinds of problems.

However, it's one of the basic tasks in statistics that you should be able to do by hand. We'll use the Z score as a way to take advantage of known properties of the standard normal curve.

$$Z = \frac{(X_i - \mu)}{\sigma}$$

$Z = 1$ (with the mean = 0 and SD = 1). Z (say "Z-score") is called the normal deviate (aka "standard normal score"; it is also called the "Z-score"); it gives us a shortcut for finding the proportion of data greater than 7 in this case).

We use the normal deviate to do a couple of things; one use is to standardize a sample of observations so that they have a mean of zero and a standard deviation of one (the **Z distribution**). The data would then said to have been **normalized**.

The second use is to make predictions about how often a particular observation is likely to be encountered. As you can imagine, this last use is very helpful for designing an experiment — if we need to see a specified difference, we can conduct a pilot study (or refer to the literature) to determine a mean and level of variability for our observation of choice, plug these back into the normal equation and predict how likely we can expect to see a particular difference. In other words, this is one way to answer that question — how many observations need I make for my experiment to be valid?

Table of normal distribution

A portion of the table of the normal curve is provided at our web site and in your workbook. For our discussions, here's another copy to look at (Fig. 6.7.2).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	
0.0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.
0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.
0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.
0.3	0.382089	0.378281	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.
0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.
0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.
0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.
0.7	0.241964	0.238852	0.235763	0.232695	0.229650	0.226627	0.223627	0.220650	0.
0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.
0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.
1.0	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.
1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.
1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.

Figure 6.7.2: Portion of the table of the normal distribution. Only values equal to or greater than $Z = 0$ are visible.

See Table 1 in the Appendix for a full version of the normal table.

We read values of Z from the first column and the first row. For $Z = 0.23$ we would scan the top row, scoot over to the fourth column, then trace to where the row and column intersect (Fig. 6.7.3); the frequency of occurrence of values at $Z = 0.23$ is 0.409046, or 40.9% (Fig. 6.7.3).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	
0.0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.
0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.
0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.
0.3	0.382089	0.378281	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.
0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.
0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.
0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.
0.7	0.241964	0.238852	0.235763	0.232695	0.229650	0.226627	0.223627	0.220650	0.
0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.
0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.
1.0	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.
1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.
1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.

Figure 6.7.3: Highlighted $Z = 0.23$ in table, frequency is 0.409046.

Z on the standard normal table is going to range between -4 and $+4$, with $Z = 0$ corresponding to 0.500. The Normal table values are symmetrical about the mean of zero.

What to make of the values of Z, from $-4, -3, \ldots +2, +3$, up to $+4$ and beyond? These are the standard deviations! Recall that using the Z score you corrected to a mean of zero (got it!), and a standard deviation of one! $Z = 2$ is twice the standard deviation; a $Z = 3$ is therefore three times the standard deviation, and so forth. The distribution is symmetrical: you get the same frequency for negative as for positive values. So on the "X" axis on a standard normal distribution, we have units of standard deviation plus

(greater) or minus (less) than the mean. In Figure 6.7.4, the area under the curve representing less than -1 standard deviations is highlighted.

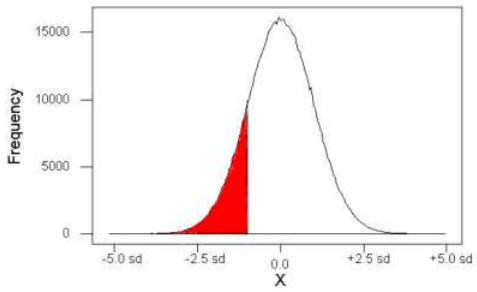


Figure 6.7.4: Plot of standard normal distribution; highlighted area under curve less than $X = -1\sigma$.

Question. How many multiples of standard deviations would you have for a Z score of $Z = 1.75$?

Answer = 1.75 times

Examples

See Table 1 in the Appendix for a full version of the normal table as you read this section.

What proportion of the data set will have values greater than 7? After applying our Z score equation, I get $Z = 1.0$, which translates to a frequency of 0.1587 or 15.87% that the observations are greater than 7.

What proportion of the data set will have values less than -7? After applying our Z score equation, I get $Z = -1.0$. Taking advantage of the symmetry argument, I just take my $Z = -1.0$ and make it positive — instead of values smaller than -1.0, we now have values greater than $+1.0$. And for $Z = 1.0$, 0.1587 or 15.87% of the observations are greater than 7, which means that 15.87% will be -7 or smaller.

What proportion of the data set will have values greater than 8? Again, apply the Z score equation. I get that for $Z = 1.5$, 0.0668 or 6.68% of the observations are greater than Z.

What proportion of the population is between 5 and 7? Draw the problem, as shown in Figure 6.7.5, where the subset of the population between 5 and 7 is colored red.

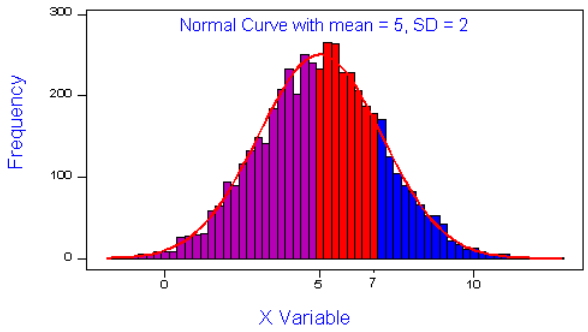


Figure 6.7.5: Proportion of the population between 5 and 7 is red (sorry about all of the colors — I kind of went crazy).

Worked problem

$1 - (\text{proportion beyond } 7) - (\text{proportion less than } 5)$

$1 - (0.1587) - (\text{proportion less than } 5)$

And the proportion less than 5?

Use the Z-score equation again. Now we find that $Z = 0$ and look up this Z-value in the table, which shows a 0.5 proportion or 50.0%.

Therefore, the proportion between 5 and 7 equals

$1 - 0.1587 - 0.50 = 0.3413$

Answer = 34.13% of the observations are between 5 and 7 when $\mu = 5$ and $\sigma = 2$.

Questions

1. Repeat the worked problem, but this time, find the proportion

- between 2 and 6.
- between 3 and 5.
- less than 5.
- greater than 7.

6.8: Moments

Introduction

Moments are used to describe the shape of a distribution. For those of you who remember your calculus, moments were discussed as a method to find the center of mass, or balancing point (Herman and Strang 2018). For distributions, the center and shape moments follow from the **expected value** of the probability function.

> **Note:**
>
> Expected value of a statistic is calculated by multiplying the likelihood of each possible outcome in a sample space by that outcome, then adding up all of those product values. From probability theory it is the weighted average of the outcomes of a random variable. A simpler way to think of the expected value is that if one were to guess the height of a person, the expected value is the average height of the population from which the person would be selected.

Four moments apply for describing the shape of a distribution. The 1st moment describes the middle, the 2nd describes the spread from the middle, the 3rd describes symmetry about the middle, and the 4th describes the shape, whether peaked and sharp, or **leptokurtic**, or broad and flattened, or **platykurtic**.

Equations for the moments

Over the years, several equations have been proposed to estimate skewness and kurtosis. The above formulas are just one example from the list (Joanes and Gill 1998).

Pearson's standardized moments:

$$\frac{\mu_n}{\sigma^n} \equiv \frac{E[X-\mu]^n}{\sigma^2}$$

where E is the expected value of the random variable. The expected value concept follows from rules of probability — basically, the average of a large number, n, of X.

Four moments can be used to describe the shape of a distribution.

1st moment, μ (mean): population mean, 3.2 – Measures of Central Tendency

2nd moment, σ^2 (variance): population variance, 3.3 – Measures of dispersion

3rd moment, $\bar{\mu}_3$ (skew):

$$\frac{\mu_3}{\sigma^3} = \frac{\sqrt{n(n-1)}}{n-2}\left[\frac{\frac{1}{n}\sum(X-\mu)^3}{\left(\frac{1}{n}\sum(X-\mu)^2\right)^{\frac{3}{2}}}\right]$$

4th moment, $\bar{\mu}_4$ (kurtosis):

$$\frac{\mu_4}{\sigma^4} = 3\sigma^4 = \frac{\mu_4}{\sigma_2^2} - 3$$

Estimating moments in R and R Commander

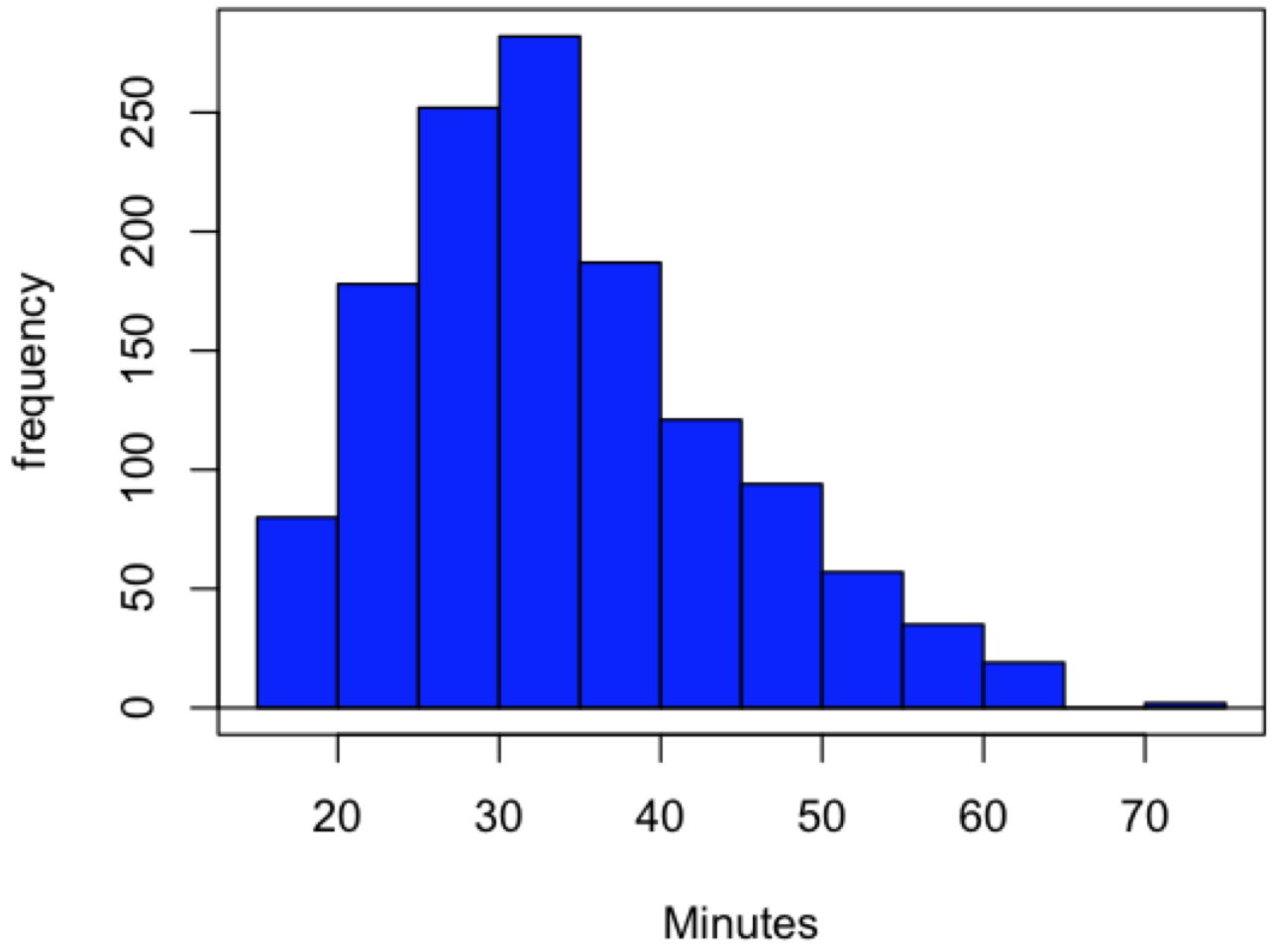


Figure 6.8.1: Histogram of finishing times in minutes for 1307 runners at the 2016 Banana 5K race.

In R Commander, we select **Statistics → Summaries → Numerical summaries…**, which brings up a popup menu. First, select the variable, in this case Minutes, from the Data tab (not shown). Next, click on Statistics tab to choose options (Fig. 6.8.2).

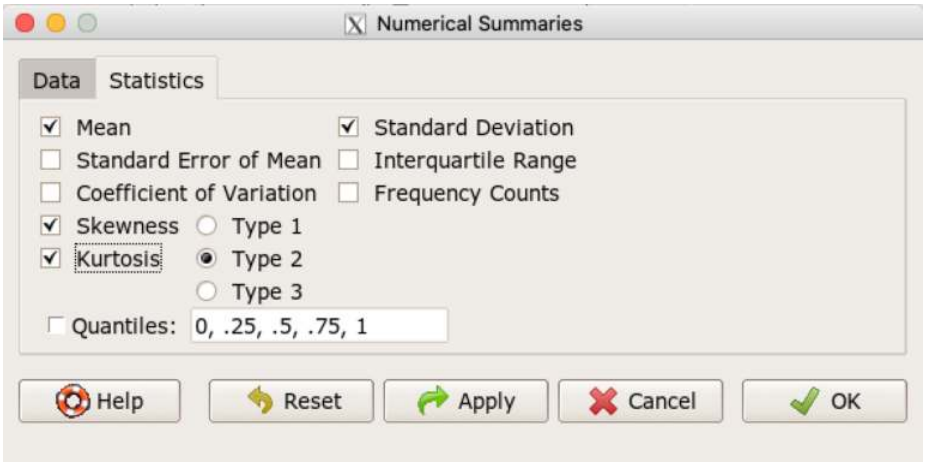


Figure 6.8.2: Numerical Summaries menu in Rcmdr Statistics.

For estimates of the moments, check Mean, Standard Deviation, Skewness, and Kurtosis. Note that Rcmdr gives you the choice among three different Types of skewness and kurtosis. Type 1 include the equations provided on this page, corresponding to definitions dating back to the 1940s. Type 2 is the default and corresponds to equations used by other professional statistics

package (SAS, SPSS). For large sample size, the different types will tend to agree. Caution applies to smaller data sets — the different types may disagree (Joanes and Gill 1998).

Large sample size, n = 1307

Type 1

```
    mean       sd  skewness    kurtosis    n
34.42999 10.31437 0.6159258 -0.01593882 1307
```

Type 2

```
    mean       sd  skewness    kurtosis    n
34.42999 10.31437 0.6166337 -0.01139521 1307
```

Type 3

```
    mean       sd skewness    kurtosis    n
34.42999 10.31437 0.615219 -0.02050335 1307
```

Small sample size

To test the claim about sample size and the moment statistics, draw a random sample of 30 from the larger data set. Sample without replacement:

```
sample.banana <- data.frame(sample(banana5K$Minutes, 30, replace = FALSE))
```

I forgot to specify a new variable name, so R used the whole command as the variable name. I could go back and fix my function call, or simply rename the variable as follows:

```
names(sample.banana)[c(1)] <- c("Minutes")
```

The random sample yielded a distribution (Fig. 6.8.3).

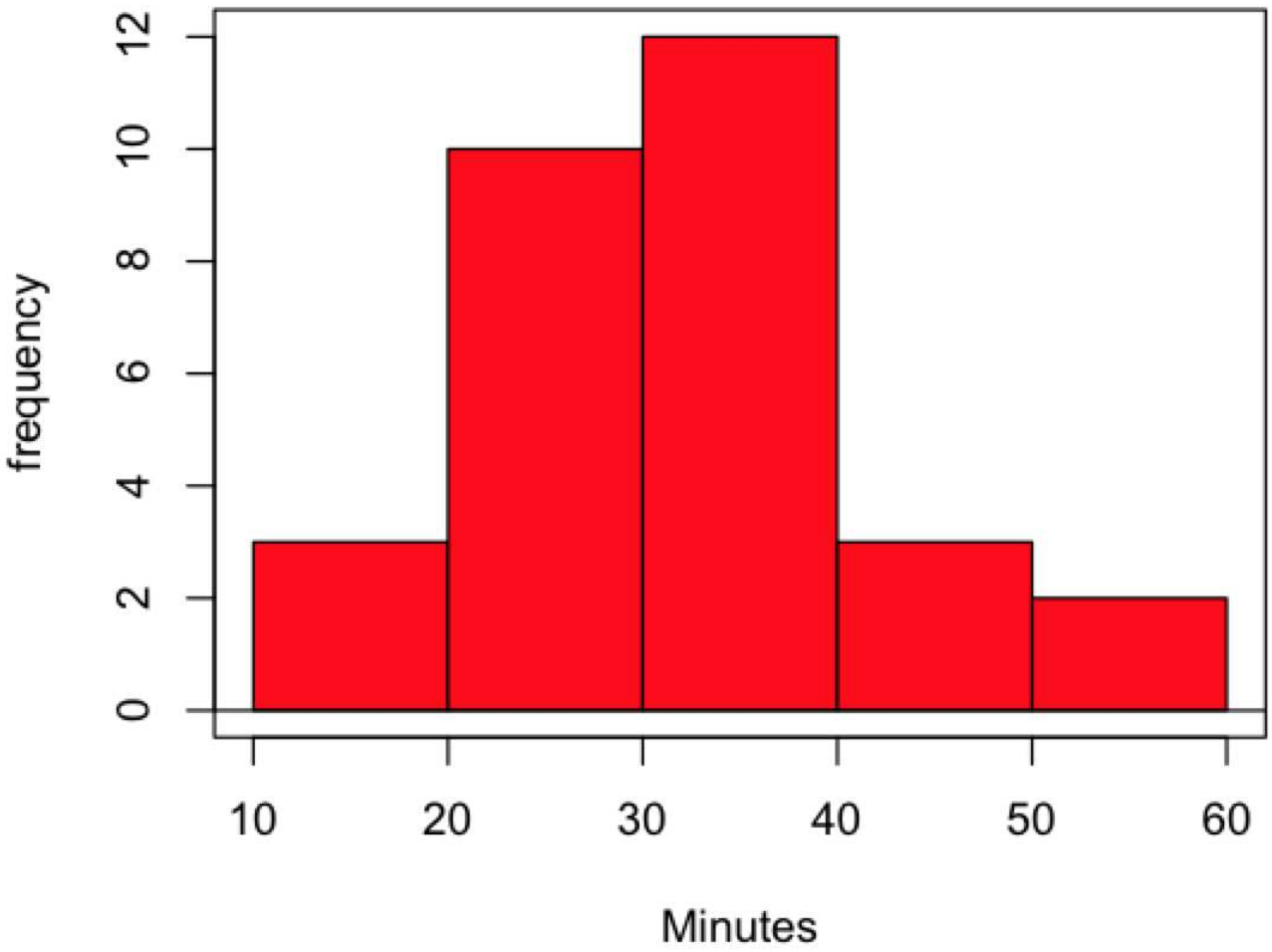


Figure 6.8.3: Histogram of finishing times in minutes for random sample of 30 drawn from 1307 runners at 2016 Banana 5K race.

Repeat Numerical summaries on small data set, n = 30

Type 1:

```
    mean       sd  skewness  kurtosis  n
33.16667 10.00373 0.5538637 0.5024438 30
```

Type 2:

```
    mean       sd  skewness  kurtosis  n
33.16667 10.00373 0.5834511 0.8276415 30
```

Type 3:

```
    mean       sd  skewness  kurtosis  n
33.16667 10.00373 0.5264025 0.2728392 30
```

Conclusion: We can compare consistency of the estimators by calculating coefficient of variation. The three types of skewness estimators differed by only 1% and 5% for large and small sample size, respectively. In contrast, the three types of kurtosis estimators differed by 29% and 52% for large and small sample size, respectively.

Questions

1. Explore the consistency of skewness and kurtosis estimates by calculating and comparing coefficient of variation estimates. R Commander provide a nice way to draw randomly from various defined distributions. Draw two data sets of 15 (small) and 1000 (large), from the chi-square distribution (1 degree of freedom) and a minimum of one other continuous distribution.

Example, draw random sample of 1000 from chi-square distribution. Rcmdr: Distributions → Continuous distributions → Chi-squared distribution → Sample from chi-squared distribution…

Enter name for the variable, enter degrees of freedom (e.g., 1), number of samples (e.g., 1000), and number of observations (variables, columns). Leave Sample means checked under data sets.

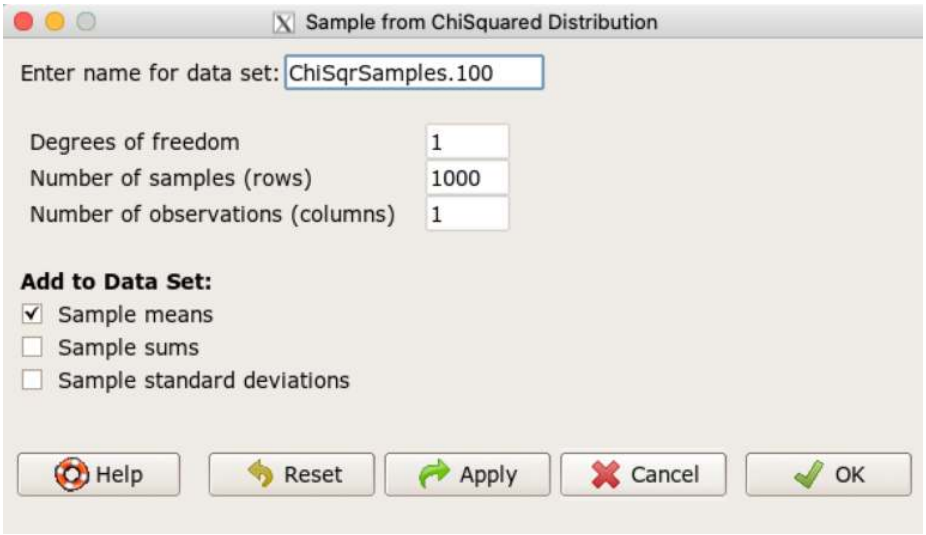


Figure 6.8.4: Sample from Chi Squared Distribution menu in Rcmdr.

This results in a new data set. Get “moments” from Numerical summaries and calculate coefficient of variations. Which moments have the most consistency regardless of the kind of distribution.

2. Make histograms for each of your created data sets. Describe what you see about the shape of the plotted distributions.

6.9: Chi-square distribution

As noted earlier, the **normal deviate** or **Z score** can be viewed as randomly sampled from the **standard normal distribution**. The chi-square distribution describes the probability distribution of the squared standardized normal deviates with **degrees of freedom**, df, equal to the number of samples taken. (The number of independent pieces of information needed to calculate the estimate, see Ch. 8.) We will use the chi-square distribution to test statistical significance of categorical variables in goodness of fit tests and contingency table problems.

The equation of the chi-square is

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(f_i - \hat{f}_i\right)^2}{\hat{f}_i}$$

where k is the number of groups or categories, from 1 to k, and f_i is the **observed frequency** and $\hat{f}_i$ is the **expected frequency** for the k^{th} category. We call the result of this calculation the chi-square test statistic. We evaluate how often that value or greater of a test statistic will occur by applying the chi-square distribution function. Graphs to show chi-square distribution for degrees of freedom equal to 1 through 5, 10, and 20 (Fig. 6.9.1).

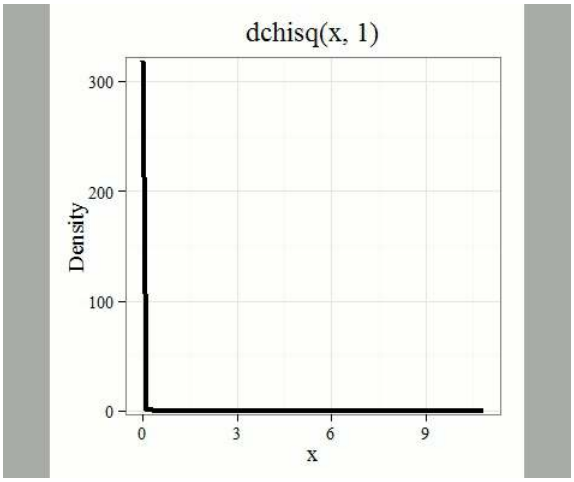


Figure 6.9.1: Animated GIF of plots of chi-square distribution over a range of degrees of freedom.

Note that the distribution is asymmetric, particularly at low **degrees of freedom**. Thus, tests using the chi-square are one-tailed (Fig. 6.9.2).

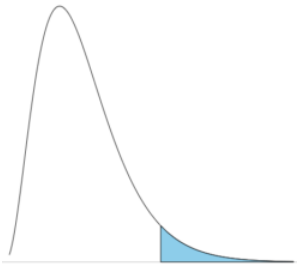


Figure 6.9.2: The test of the chi-square is typically one-tailed. In this case, the highlighted area shows the probability of values greater than the critical value.

By convention in the Null Hypothesis Significance Testing protocol (NHST), we compare the test statistic to a **critical value**. The critical value is defined as the value of the test statistic — the cutoff boundary between **statistical significance** and insignificance — that occurs at the Type I error rate, which is typically set to 5%. The interpretation of the result is as follows: after calculating a test statistic, we can judge significance of the results relative to the null hypothesis expectation. If our test statistics is greater than the critical value, then the **p-value** of our results are less than 5% (R will report an exact p-value for the test statistic). You are not expected to be able to follow this logic just yet — rather, we teach it now as a sort of mechanical understanding to develop in the NHST tradition. The justification for this approach to testing of statistical significance is developed in Chapter 8. A portion of the critical values of the chi-square distribution are shown in Figure 6.9.3.

α(1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
DF/1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.108	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.730
4	5.385	7.779	9.488	11.143	13.277	14.860	16.424	18.467	19.997
5	6.626	9.236	11.070	12.833	15.086	16.750	18.386	20.515	22.105
6	7.841	10.645	12.592	14.449	16.812	18.548	20.249	22.458	24.103
7	9.037	12.017	14.067	16.013	18.475	20.278	22.040	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.868
9	11.389	14.684	16.919	19.023	21.666	23.589	25.462	27.877	29.666
10	12.549	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.420

Figure 6.9.3: Portion of the table of some critical values of chi-square distribution, one tailed (right-tailed or "upper" portion of distribution).

See Appendix for a complete chi-square table.

Example

Professor Hermon Bumpus of Brown University in Providence, Rhode Island, received 136 House Sparrows (*Passer domesticus*) after a severe winter storm 1 February 1898. The birds were collected from the ground; 72 of the birds survived, 64 did not (Table 6.9.1). Bumpus made several measures of morphology on the birds and the data set has served as a classical example of Natural Selection (Chicago Field Museum). We'll look at this data set when we introduce Linear Regression.

Table 6.9.1. Survival statistics of Bumpus House sparrows.

	Yes	No
Female	21	28
Male	51	36

Was there a survival difference between male and female House Sparrows? This is a classic contingency table analysis, something we will at length in Chapter 9. For now, we report the Chi-square test statistic for this test was 3.1264 and the test had one degree of freedom. What is the critical value of the chi-square distribution at 5% and one degree of freedom?. Typically we would simply use R to look this up

```
qchisq(c(0.05), df=1, lower.tail=FALSE)
```

But we can also get this from the table of critical values (Fig. 6.9.4). Simply select the row based on the degrees of freedom for the test then scan to the column with the appropriate significance level, again, typically 5% (0.05).

α(1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
DF/1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.108	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.730
4	5.385	7.779	9.488	11.143	13.277	14.860	16.424	18.467	19.997
5	6.626	9.236	11.070	12.833	15.086	16.750	18.386	20.515	22.105
6	7.841	10.645	12.592	14.449	16.812	18.548	20.249	22.458	24.103
7	9.037	12.017	14.067	16.013	18.475	20.278	22.040	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.868
9	11.389	14.684	16.919	19.023	21.666	23.589	25.462	27.877	29.666
10	12.549	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.420

Figure 6.9.4: Portion of the chi-square distribution which shows how to find critical value of the chi-square distribution.

For 1 degree of freedom at 5% significance, the critical value is 3.841. Back to our hypothesis: Did male and female survival differ in the Bumpus data set? Following the NHST logic, if the test statistic value (e.g., 3.1264) is greater than the critical value (3.841), then we would reject the null hypothesis. For this example, we would conclude no statistical difference between male and female survival because the test statistic was smaller than the critical value. How likely are these results due to chance? That's where the p-value comes in. Our test statistic value falls between 5% and 10% ($2.706 < 3.1264 < 3.841$). In order to get the actual p-value of our test statistic we would need to use R.

R code

Given a chi-square test statistic, you can use R to calculate the probability of that value against the null hypothesis. At the R prompt, enter:

```
pchisq(c(3.1264), df=1, lower.tail=FALSE)
```

The R output is

```
[1] 0.07703368
```

Because we are using R Commander, simply select the command by following the menu options.

Rcmdr: Distributions → Continuous distributions → Chi-squared distribution → Chi-squared probabilities …

Enter the chi-square value and degrees of freedom (Fig. 6.9.5).

Figure 6.9.5: Screenshot of input box in Rcmdr for Chi-square probability values.

Questions

1. What happens to the shape of the chi-square distribution as degrees of freedom are increased from 1 to 5 to 20 to 100?

Be able to answer the following questions using the Chi-square table or using Rcmdr:

2. For probability $\alpha = 5\%$, what is the critical value of the chi-square distribution (upper tail)?
3. The value of the chi-square test statistic is given as 12. With 3 degrees of freedom, what is the approximate probability of this value or greater from the chi-square distribution?

6.10: t-distribution

Introduction

Student's t distribution is a sampling distribution where values are sampled from a normal distributed population, but σ, the standard deviation, and μ, the mean of the population, are not known. When sample size is large and we know the standard deviation, we would use the Z-score to evaluate probabilities of the sample mean. The t-distribution applies when σ is not known and the sample size is small (e.g., less than 30, per **rule of thirty**).

> Note:
>
> According to Wikipedia and sources therein, Student was the pseudonym of William Sealy Gosset, who came up with the t-test and t-distribution.

The equation of the **t-test** is

$$t = \frac{(\bar{X} - \mu)}{s_{\bar{X}}}$$

where the difference between $\bar{X}$ the sample mean, and μ, the population mean, is divided by the **standard error of the mean**, $s_{\bar{X}}$, defined in Chapter 3.2 and again in Chapter 3.3. This formulation of the t-test is called the **one sample t-test** (Chapter 8.5). We call the result of this calculation the **test statistic** for t. We evaluate how often that value or greater of a test statistic will occur by applying the t distribution function.

There are many t-distributions, actually, one for every degree of freedom. Like the normal distribution, the t distribution is symmetrical about a mean of zero. But it is stacked up (leptokurtic) around the middle at low degrees of freedom. As degrees of freedom increase, the t distribution spreads and becomes increasingly like the normal distribution.

Relationship between t distribution and standard normal curve

First, here is our standard normal plot, mean = 0, standard deviation = 1

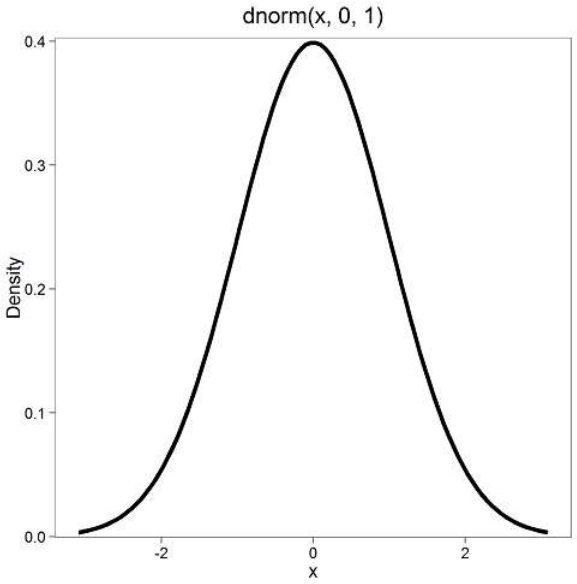


Figure 6.10.1: Plot of standard normal distribution.

Next, here's the t-distribution for five degrees of freedom.

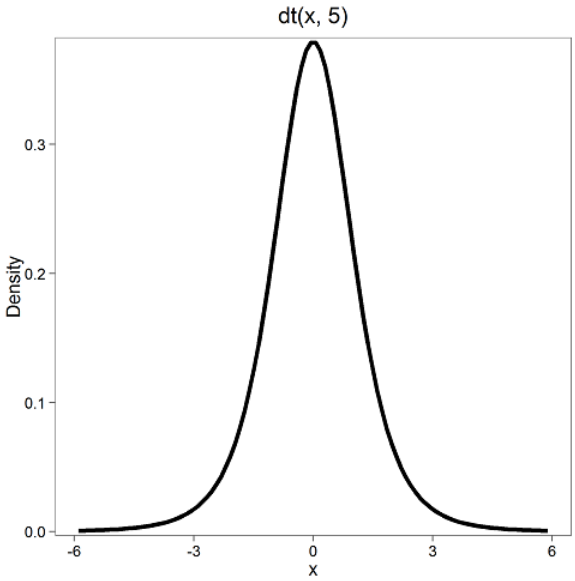


Figure 6.10.2: Plot of t-distribution for 5 degrees of freedom.

Lets see what happens to the shape of the t-distribution as we increase the degrees of freedom from $df = 5, 10, 20, 50, 1000, 10000$The last graphic is the standard normal curve again.

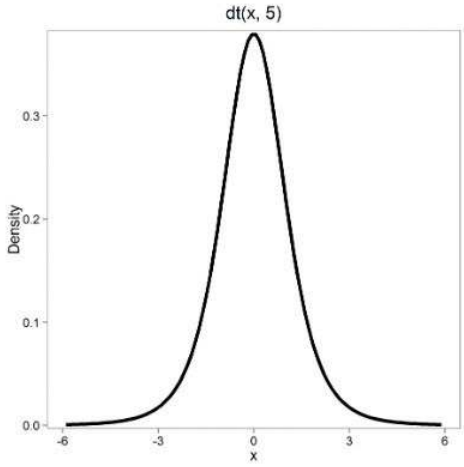


Figure 6.10.3: Animated GIF of t-distribution plots, from $df = 5$ to $10,000$ plus standard normal curve.

By convention in the Null Hypothesis Significance Testing protocol (NHST), we compare the test statistic to a critical value. The critical value is defined as the value of the test statistic that occurs at the Type I error rate, which is typically set to 5%. We introduced logic of NHST approach in Chapter 6.9 with the chi-square distribution. Again ,this is just an introduction; we teach it now as a sort of mechanical understanding to develop. The justification for this approach to testing of statistical significance is developed in Chapter 8.

Table 6.10.1. Critical values of the t-distribution for $df = 1, \ldots, 5$, one tail (upper)

df	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	6.314	12.706	31.820
2	2.920	4.303	6.965
3	2.353	3.182	4.541
4	2.132	2.776	3.747
5	2.015	2.571	3.365

See Appendix A.4 for a complete table of t-distribution.

Questions

1. What happens to the shape of the t distribution as degrees of freedom are increased from 1 to 5 to 20 to 100?

Be able to answer these questions using the t table in Appendix 20.4, or using Rcmdr:

2. For probability $\alpha = 5\%$, what is the critical value of the t distribution (upper tail) for 1 degree of freedom? For $df = 5$? For $df = 20$? For $df = 30$?
3. The value of the t test statistic is given as 12. With 3 degrees of freedom, what is the approximate probability of this value or greater from the t distribution?

6.11: F-distribution

Introduction

The ***F* distribution** is the probability distribution associated with the ***F* statistic** and named in honor of R. A. Fisher. The F distribution is used as the null distribution of the ANOVA test statistic. The F distribution is the ratio of two **chi-square distributions**, with degrees of freedom v_1 and v_2 for numerator and denominator, respectively.

We can for illustration purposes define the F statistic as a ratio of two variances,

$$F = \frac{s_2^2}{s_1^2}$$

The F statistic has two sets of degrees of freedom, one for the numerator and one for the denominator. The actual formula for the F distribution is quite complicated and in general we don't use the F distribution in a way that involves parameter estimation. Rather, it is used in evaluating the statistical significance of the F statistic. Therefore, we produce but a few graphs and a table of critical values to illustrate the distribution.

We call the result of this calculation the F test statistic. We evaluate how often that value or greater of a test statistic will occur by applying the F distribution function. A few graphs to get a sense of what the distribution looks like for varying v_1, v_2 values held to ten degrees of freedom (Fig. 6.11.1).

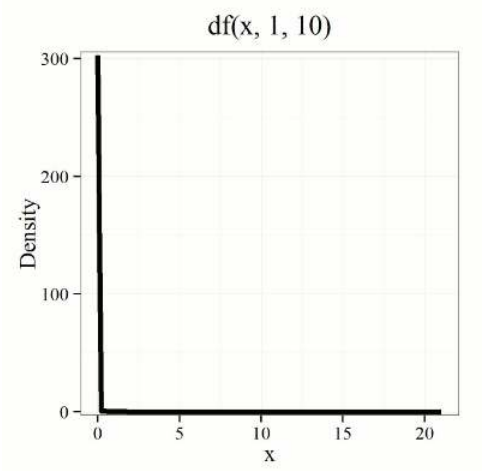


Figure 6.11.1: Animated GIF plot of F distribution value for range of degrees of freedom.

By convention in the Null Hypothesis Significance Testing protocol (NHST), we compare the test statistic to a critical value. The critical value is defined as the value of the test statistic that occurs at the Type I error rate, which is typically set to 5%, per our presentations in Chapter 6.7, 6.9, and 6.10. The justification for NHST approach to testing of statistical significance is developed in Chapter 8.

Table 6.11.1. Critical values of the F distribution, one tail (upper), degrees of freedom $v_1 = 1$ through $v_1 = 4$, $v_2 = 10$

\(v_{1\)	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	4.964	6.937	10.044
2	4.103	5.456	7.559
3	3.708	4.826	6.552
4	3.478	4.468	5.994

For the complete F table see Appendix A.5.

χ^2, t and F distributions are related

χ^2, t and F distributions are all distributions indexed by their degrees of freedom. With some algebra, these three distributions can be shown to be related to each other. The probabilities tabled in the chi-squared are part of the F-distribution.

Some interesting relationships between the F distribution and other distributions can be shown. By definition we claimed that the F distribution is built on ratio of chi-square distributions, so that should indicate to you the relationship between the two kinds of continuous probability distributions. However, one can also show relationships to other distributions for the F distribution. For example, for the case of $v_1 = 1$ and v_2 = any value, then $F_{1,v_2} = t^2$, where t refers to the t distribution.

Questions

1. What happens to the shape of the F distribution as degrees of freedom are increased from 1 to 5 to 20 to 100?
2. In Rcmdr, which option do you select to get the critical value for $df = 1$ and $df = 20$ at $\alpha = 5\%$?
   A. F quantiles
   B. F probabilities
   C. Plot of F distribution
   D. Sample from F distribution

Be able to answer these questions using the F table, Appendix A.5, or using Rcmdr

3. For probability $\alpha = 5\%$, and numerator degrees of freedom equal to 1, what is the critical value of the F distribution (upper tail) for 1 degree of freedom? For $df = 5$? For $df = 20$? For $df = 30$?
4. The value of the F test statistic is given as 12. With 3 degrees of freedom for the numerator, and ten degrees of freedom for the denominator, what is the approximate probability of this value, or greater from the F distribution?

6.12: Chapter 6 References and Suggested Readings

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604).

Baker, D., Lidster, K., Sottomayor, A., Amor, S. (2014). Two years later: Journals are not yet enforcing the ARRIVE Guidelines on reporting standards for pre-clinical animal studies. *PLoS Biology* 12:e1001756

Bewick, V., Cheek, L., Ball, J. (2002). Statistics review 8: Qualitative data – tests of association. *Clinical Care* 8:46-58.

Borge, K. S., Tønnessen, R., Nødtvedt, A., Indrebø, A. (2011). Litter size at birth in purebred dogs—A retrospective study of 224 breeds. *Theriogenology* 75:911–919

Bumpus, H. C. (1898). The variations and mutations of the introduced sparrow, *Passer domesticus*. *Biological Lectures Delivered at the Marine Biological Laboratory of Woods Hole*, 1896-1897, pp. 1-15. (link to article at Biodiversity Heritage Library)

Bumpus, H. C. (1899). The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. *Biology Lectures delivered at the Marine Biology Laboratory, Woods Hole*, 1896-7, 209-226 (link to volume Google Books).

Di Traglia, F., Nolesini, T., Solari, L., Ciampalini, A., Frodella, W., Steri, D., … & Galardi, E. (2018). Lava delta deformation as a proxy for submarine slope instability. *Earth and Planetary Science Letters*, *488*, 46-58.

Diaconis, P. (1978). Statistical problems in ESP research. *Science*, *201*(4351), 131-136.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. Simon & Schuster

Drink up: Coffee is good for you, research shows. (n.d.). Retrieved from https://www.nbcnews.com/health/health-news/coffee-good-you-more-science-shows-n888356

Herman, E., & Strang, G. (2018). Calculus Volume 1, 2, and 3. OpenStax

Husak, J. F., Fox, S. F., Lovern, M. B., & Bussche, R. A. V. D. (2006). Faster lizards sire more offspring: sexual selection on whole-animal performance. Evolution, 60(10), 2122-2130.

International Shark Attack file at Florida Museum of Natural History, https://www.floridamuseum.ufl.edu/shark-attacks/

Jeffrey, R. (1995). Probabalistic thinking. Online book at http://www.princeton.edu/~bayesway/ProbThink/

Jasieński, M., & Bazzaz, F. A. (1999). The fallacy of ratios and the testability of models in biology. *Oikos*, 321-326.

Karp, N. A., Segonds-Pichon, A., Gerdin, A. K. B., Ramírez-Solis, R., & White, J. K. (2012). The fallacy of ratio correction to address confounding factors. *Laboratory animals*, 46(3), 245-252.

Langley (2009). Human Fatalities Resulting From Dog Attacks in the United States, 1979–2005. *Wilderness and Environmental Medicine* 20(1):19–25.

Leemis, L. M., McQuestion, J. T. (2008). Univariate distribution relationships. *The American Statistician* 62:45-53.

Liermann, M., Steel, A., Rosing, M., & Guttorp, P. (2004). Random denominators and the analysis of ratio data. *Environmental and ecological statistics*, *11*(1), 55-71.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105*(1), 156.

Mincarelli, L., Tiano, L., Craft, J., Marcheggiani, F., & Vischetti, C. (2019). Evaluation of gene expression of different molecular biomarkers of stress response as an effect of copper exposure on the earthworm EIsenia Andrei. *Ecotoxicology*, 28(8), 938-948.

Moretti, T. R., Moreno, L. I., Smerick, J. B., Pignone, M. L., Hizon, R., Buckleton, J. S., … & Onorato, A. J. (2016). Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. *Forensic Science International: Genetics*, *25*, 175-181.

Murray, D. B., & Teare, S. W. (1993). Probability of a tossed coin landing on edge. *Physical Review E*, *48*(4), 2547.

National Research Council (1996). Chapter 4. Population genetics. In *The Evaluation of Forensic DNA Evidence*. Washington, D.C.: The National Academies Press. doi: https://doi.org/10.17226/5141

Nee, S., Colegrave, N., West, S. A., & Grafen, A. (2005). The illusion of invariant quantities in life histories. *Science*, 309(5738), 1236-1239.

Norman, G., Streiner, D. (2003). Describing data, Chapter 2, *PDQ statistics*. B C Decker

Packard, G. C., & Boardman, T. J. (1988). The misuse of ratios, indices, and percentages in ecophysiological research. *Physiological Zoology*, 61(1), 1-9.

Pigliucci, M. (2010). *Nonsense on stilts: How to tell science from bunk*. University of Chicago Press.

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education*, *5*(4), 541-542.

CHAPTER OVERVIEW

7: Probability and Risk Analysis

Introduction

The public health applications of **epidemiology,** the branch of medicine concerned with identifying patterns and potential causes of disease and health in populations, were every day in the news during the Covid-19 pandemic. From contact tracing to reproductive rate of the SARS-Cov-2 virus to numbers of hospital beds and nurses available in ICU units across the country, to the discussions and debates over how the virus is spread, no doubt you have learned much about the critical role epidemiology continues to play.

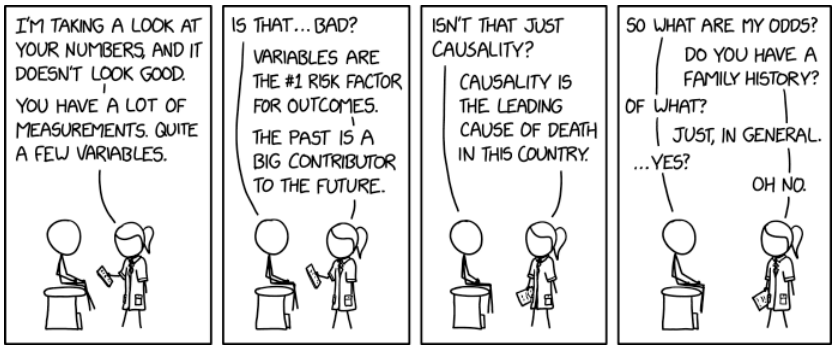


Figure 7.1: "Health data," https://xkcd.com/2620/

This chapter is about probability and will introduce you to **risk analysis** (Fig. 7.1), used to "… characterize the nature and magnitude of risks to human health for various populations…", a foundational topic in biostatistics and epidemiology. The `epiR` package will be introduced and code examples provided for **descriptive epidemiology** and again for **statistical inference** (Chapter 9).

7.1: Epidemiology definitions

Introduction

This sub-chapter may lack for drama, but let's start by providing a list of key terms with definitions as you start our introduction to epidemiology.

Definitions

Absolute risk: The probability that a specified event will occur in a specified population. See Ch. 7.4 – Epidemiology: Relative risk and absolute risk, explained

Absolute risk reduction (ARR): the decrease in risk of an event in an exposed (treatment) group compared to an unexposed (control) group. Also called the **risk difference**. $CER - EER$, see Contingency table. See Ch. 7.4 – Epidemiology: Relative risk and absolute risk, explained

Contingency table, also called cross tabulation or crosstab, is a display of counts of variables in a matrix format. In epidemiology, rows of contingency table represent treatment or exposure groups, and columns represent outcomes.

Table 7.1.1. A 2 × 2 contingency table.

	Outcome	
	Yes	**No**
Treatment or exposed group	**a**	**b**
Control or nonexposed group	**c**	**d**

The 2 × 2 contingency table is referred to frequently in this chapter and again in Chapter 9.

R code

```
a = 4; b = 46; c = 5; d = 45

Table1 <- matrix(c(a,b,c,d), 2, 2, byrow=TRUE, dimnames = list(c("Treatment", "Contro
```

R output

```
          Yes  No
Treatment   4  46
Control     5  45
```

Control event rate (CER): How often an event occurs in the control group. $\frac{c}{c+d}$, see Contingency table

Diagnosis: identification of the nature of a disease or condition.

Event: From probability theory, an event is a set of outcomes to which a probability is assigned.

Experimental event rate (EER): How often an event occurs in the treatment group. $\frac{a}{a+b}$, see Contingency table

Hazard: anything that can cause harm

Incidence: the number of newly diagnosed individuals in a population having a condition, disease or other characteristic. Compare to prevalence.

Negative predictive value of a test (**NPV**), defined as the probability that a negative test result identifies a person who truly does not have the disease. Calculated as the total number of individuals without the disease divided by the total that tested negative. $NPV = \frac{TN}{TN+FN}$

Number needed to treat (NNT): the inverse of the absolute risk reduction. $NNT = \frac{1}{ARR}$. See Ch. 7.4 – Epidemiology: Relative risk and absolute risk, explained

Odds, the ratio (OR) of two probabilities: the probability of getting a one on throwing a dice is 1, and the probability of not getting a one is 5; therefore the odds of getting a one are 1 to 5. $OR = \frac{a \cdot d}{b \cdot c}$. See Ch. 7.5 – Odds ratio

Per capita rate, Latin phrase, for each head, meaning per person.

Positive predictive value of a test (**PPV**), defined as the probability that a positive test result identifies a person who truly has the disease. Calculated as the total number of individuals with the disease divided by the total that tested positive. $PPV = \frac{TP}{TP+FP}$

Posttest probability refers to the probability that the patient has the disease after the results of the test are known.

Pretest probability is the prevalence of the disease, i.e., the chance that the a randomly selected person from the population has the disease.

Prevalence: The proportion of individuals in a population having a condition, disease, or characteristic. Compare to incidence.

Prognosis: how a disease plays out.

Relative risk: Ratio of the risk of an event among those exposed to the risk factor to the risk among those not exposed to the risk factor. See Ch. 7.4 – Epidemiology: Relative risk and absolute risk, explained

Relative risk reduction (RRR): is a measure calculated by dividing the absolute risk reduction by the control event rate. See Ch. 7.4 – Epidemiology: Relative risk and absolute risk, explained

Risk: Probability of an event. Risk is not restricted to just bad events, but refers to the uncertainty of a particular event (e.g., the risk that a child will be born male seems a melodramatic statement, but it is accurate as far as this definition goes).

Therapy: treatment intended to treat, relieve, or cure a disorder or condition.

Questions

1. Compare and contrast ARR and RRR.
2. What's the difference between event, hazard, and risk?
3. What's the difference between incidence and prevalence?
4. What's the difference between diagnosis and prognosis?
5. What's the implication of a NNT greater than 100 in terms of the utility of a proposed therapy or treatment?

7.2: Epidemiology basics

Introduction

To introduce **conditional probability**, I have elected to push you into epidemiology and risk analysis. We introduced epidemiology definitions and here, we build on basic terminology of epidemiology. Epidemiology is the study of the causes and distribution of health-related events in a population. It is has been called the basic science of public health (p. 16, Decker 2008).

Prevalence rate

Prevalence of a disease (or condition), is defined as the proportion of the population that has the disease (condition) at a point or duration in time. The prevalence statistic is the ratio of the number of existing cases divided by the total population.

For example, let's say we're interested in the prevalence of Type 2 diabetes in Hawai`i. The 2020 populations was about 1.4 million. A survey is conducted on a random of 1000 individuals and 80 were reported to have Type 2 diabetes. What is the estimated prevalence of Type 2 diabetes in Hawai`i?

Start from perspective — what if we actually have something close to a census count? For the actual estimates, see the report at the Hawaii State DOH website. With every estimate of a statistic, we need a confidence interval (CI). An approximate (good for large samples) formula for the **95% CI of prevalence** is

$$95\%\, CI = \hat{p} \pm z\sqrt{\frac{1}{N}\hat{p}\,(1-\hat{p})}$$

where $\hat{p}$ is the 2010 prevalence in the population (8.3%), N is the population size (1,360,301), and z is the **standard normal probability**. For a 95% CI, then we want $z_{95\%}$.

If you recall, we can get this from our standard normal probability table, or directly from R. We want to know z that is in the $\pm 2.5\%$ **tails of the distribution** (that's $0.05 \div 2$, see Ch 8.4 – Tails of a test). Our R code then is

```
qnorm(c(0.025), mean=0, sd=1, lower.tail=FALSE)
```

which returns

```
[1] 1.959964
```

You should confirm that setting `lower-tail = TRUE` yields -1.96 (rounded).

Alternately, if using **Rcmdr: Distributions → Continuous distributions → Normal distribution → Normal quantiles…** (Fig. 1)

Figure 7.2.1: R Commander popup menu for Normal Quantiles.

Thus, for our example, the 95% CI was $(8.3 - 0.046, 8.3 + 0.046)$; our confidence in our estimate of the prevalence of diabetes in Hawaii is between 8.25% and 8.35%. Again, note that for our purposes it is OK to calculate the *approximate* confidence interval, e.g. replace ± 1.96 with ± 2 (for large N, the differences are observed in the 0.001 decimal).

Prevalence: R code

Rather than census counts, more likely we have results of smaller surveys. We use the `epi.conf()` function from the `epiR` package. Code adapted from example provided in **epiR_descriptive** vignette.

Note:

Many R packages include vignettes, which, together with the package manual, is often helpful to understand what a function is intended to do and how to get the most from the function. R code to call all vignettes available for a package, e.g., `epiR` :

```
vignette(package="epiR")
```

which will return names of available vignettes. For epiR, these are

```
epiR_descriptive Descriptive epidemiology (source, html)
epiR_surveillance Disease surveillance (source, html)
epiR_measures_of_association Measures of association (source, html)
epiR_sample_size Sample size calculations (source, html)
```

To call up the vignette we need for this example, use

```
vignette("epiR_descriptive", package="epiR")
```

which brings up the page (assuming you installed the html help files during installation of base R).

```
library(epiR)
pop.Hawaii = 1.4e06
pop.Survey = 1000
type2.Survey = 80
Table.2 <- as.matrix(cbind(type2.Survey, pop.Survey))
epi.conf(Table.2, ctype = "prevalence", method = "exact", N = pop.Hawaii, design = 1,
```

R output:

```
   est    lower    upper
1    8 6.394198 9.857978
```

Thus, estimated Type 2 diabetes prevalence from the survey was 8, with 95% confidence interval of 6.4 to 9.9 cases per 100 individuals. From 2020 U.S. Census, Hawai`i population was 1,455,271.

Note:

Type 2 diabetes is one of many conditions for which prevalence is greater among Native Hawaiian and Pacific Islander populations compared to other groups in Hawai'i (Galinski et al 2016); these are called **health disparities**.

Incidence rate

Incidence of a disease (or condition) is defined as the occurrence of new cases of a disease. The simplest way to view incidence is that if everyone was followed for the same period of time, then incidence rate is the number of new cases since the start of the study divided by the total population. This is too simplistic, so we define a better metric called **person-time**. **Incidence rate (IR)** is then the ratio of the number of new cases divided by total person-time.

Again, incidence rate is an estimate. Therefore, we need a confidence interval. Assuming the population is large, then confidence interval can be calculated as

$$IR \pm z\sqrt{\frac{N}{T^2}}$$

where IR is the incidence rate, T is the person-time, N is the number of events or cases, and z again is the standard normal probability ($z = \pm 1.96$ for 95% confidence interval).

For our example, N = 3 cases, T = 236 person-days, so 95% CI is $(11.3\%, 14.1\%)$

Person-time

Person-time can be days, months, years. Person-time is best defined with an example.

Five men join a study that will last 70 days. These men were selected because they had suffered a myocardial infarction (MI) and at the start of the study, they receive the same treatment. The outcome of the study is whether or not the subjects suffer a second MI.

The results, recording the number of days that passed before each subject suffered a second MI or else the full span of the study for subjects who did not suffer a second MI, are shown below:

Subject A, 53 days
Subject B, 70 days
Subject C, 24 days
Subject D, 70 days
Subject E, 19 days

Add up the person days, $53 + 70 + 24 + 70 + 19 = 236$ person-days (p-d)

Now calculate the incidence rate:

$$1000 \cdot \frac{3 \text{ cases}}{236 \text{ p-d}} = 12.7 \text{person-days}$$

that's 3 cases (A, C, E) divided by 236 p-d = 0.0127. Multiply this by 1000 and we get our final answer for the incidence rate, 12.7 per person-days.

Incidence rate: R code

Incidence rate of myocardial infarction (MI) study. Code adapted from example provided in `epiR_descriptive vignette` .

```
ncas = 3
ntar = 236
tmp <- as.matrix(cbind(ncas, ntar))
epi.conf(tmp, ctype = "inc.rate", method = "exact", N = 1000, design = 1,
conf.level = 0.95) * 1000
```

R output:

```
          est    lower    upper
ncas 12.71186 2.621492 37.14946
```

The incidence rate of myocardial infarction was 12.7 (95% CI 2.6 to 37.2) cases per 1000 person-days.

Age-specific rates

Incidence or prevalence rates may be reported for specific age groups. For example, we can distinguish between number of live births in Hawaii in 2017, and numbers of live births by age group of the mother.

Age-adjusted rates

If populations with different age demographics, then the convention is to adjust the populations to a standard reference population with known age and other demographic properties. For example, the CDC uses the 2000 census as the standard population (CDC definitions, age adjustment, retrieved January 2023).

Questions

1. Calculate the confidence interval of type 2 prevalence in Hawai`i with the 2020 census population value. How much did it change?
2. Recalculate the confidence interval for a 99% confidence interval. Which estimate communicates greater confidence in the estimate?

7.3: Conditional probability and evidence-based medicine

Introduction

This section covers a lot of ground. We begin by addressing how probability of multiple events are calculated assuming each event is independent. The assumption of independence is then relaxed, and how to determine probability of an event happening given another event has already occurred, conditional probability is introduced. Use of **conditional probability** to interpret results of a clinical test are also introduced, along with the concept of "**evidence-based-medicine**," or EBM.

Probability and independent events

Probability distributions are mathematical descriptions of the probabilities of how often different possible outcomes occur. We also introduced basic concepts related to working with the probabilities involving more than one event.

For review, for **independent events**, you multiply the individual chance that each event occurs to get the overall probability.

Example of multiple, independent events

Figure 7.3.1: Now that's a box full of kittens. Creative Commons License, source: https://www.flickr.com/photos/83014408@N00/160490011

What is the chance of five kittens in a litter of five to be of the same sex? In feral cat colonies, siblings in a litter share the same mother, but not necessarily the same father, superfecundation. Singleton births are independent events, thus the probability of the first kitten being female is 50%; the second kitten being female, also 50%; and so on. We can multiply the independent probabilities (hence, the **multiplicative rule**), to get our answer:

```
kittens <- c(0.5, 0.5, 0.5, 0.5, 0.5)
prod(kittens)
[1] 0.03125
```

Probabilistic risk analysis

Risk analysis is the use of information to identify hazards and to estimate the risk. A more serious example. Consider the 1986 *Challenger* Space Shuttle Disaster (Hastings 2003). Among the crew killed was Ellison Onizuka, the first Asian American to fly in space (Fig. 7.3.2, first on left back row). Onizuka was born and raised on Hawai`i and graduated from Konawaena High School in 1964.

Figure 7.3.2: STS-51-L crew: (front row) Michael J. Smith, Dick Scobee, Ronald McNair; (back row) Ellison Onizuka, Christa McAuliffe, Gregory Jarvis, Judith Resnik. Image by NASA – NASA Human Space Flight Gallery, Public Domain.

The shuttle was destroyed just 73 seconds after liftoff (Fig. 7.3.3).

Figure 7.3.3: Space shuttle *Challenger* launches from launchpad 39B Kennedy Space Center, FL, at the start of STS-51-L. Hundreds of shorebirds in flight.

This next section relies on material and analysis presented in the Rogers Commission Report June 1986. NASA had estimated that the probability of one engine failure would be 1 in 100, or 0.01; two engine failures would mean the shuttle would be lost. Thus, the probability of two rockets failing at the same time was calculated as 0.01 × 0.01, which is 0.0001 or 0.01%.

NASA had planned to fly the fleet of shuttles 100 times per year, which would translate to a shuttle failure once in 100 years. The Challenger launch on January 28, 1986, represented only the 25th flight of the shuttle fleet.

One difference on launch day was that the air temperature at Cape Canaveral was quite low for that time of year, as low as 22°F overnight.

Attention was pointed at the large O-rings in the boosters (engines). In all, there were six of these O-rings. Testing suggested that, at the colder air temperatures, the chance that one of the rings would fail was 0.023. Thus, the chance of success was only 0.977. Assuming independence, what is the chance that the shuttle would experience O-ring failure?

```
shuttle <- c(0.977, 0.977, 0.977, 0.977, 0.977, 0.977)
#probability of success then was
prod(shuttle)
[1] 0.869
```

```
#and therefore probability of failure was
1 - prod(shuttle)
[1] 0.1303042
```

Conditional probability of non-independent events

But in many other cases, independence of events cannot be assumed. The probability of an event given that another event has occurred is referred to as **conditional probability**. Conditional probability is used extensively to convey risk. We've touched on some of these examples already:

- the risk of subsequent coronary events given high cholesterol;
- the risk of lung cancer given that a person smokes tobacco;
- the risk of mortality from breast cancer given that regular mammography screening was conducted.

There are many, many examples in medicine, insurance, you name it. It is even an important concept that judges and lawyers need to be able to handle (e.g., Berry 2008).

A now famous example of conditional probability in the legal arena came from arguments over the chance that a husband or partner who beats his wife will subsequently murder her — this was an argument raised by the prosecution during pre-trial in the 1995 OJ Simpson trial (*The People of the State of California v. Orenthal James Simpson*), and successfully argued by O.J. Simpson's attorneys… the judge ruled in favor of the defense, and evidence of OJ Simpson's prior abuse were not included in trial. Gigerenzer (2002) and others have called this reverse **Prosecutor's Fallacy**, where the more typical scenario is that the prosecution provides a list of probabilities about characteristics of the defendant, leaving the jury to conclude that no one else could have possibly fit the description. In the OJ Simpson case, the argument went something like this. From the CDC we find that an estimated 1.3 million women are abused each year by their partner or former partner; each year about 1000 women are murdered. One thousand divided by 1.3 million is a small number, so even when there is abuse, the argument goes, 99% of the time there is not murder. The Simpson judge ruled in favor of the defense and much of the evidence of abuse was excluded.

Something is missing from the defense's argument. Nicole Simpson did not belong to a population of battered women — she belonged to the population of murdered women. When we ask, if a woman is murdered, what is the chance that she knew her murderer, we find that more than 55% knew their murderer — and of that 55%, 93% were killed by a current partner. The key is that Nicole Simpson (and Ron Goldman) was murdered and OJ Simpson was an ex-partner who had been guilty of assault against Nicole Simpson. Now, it goes from an impossibly small chance, to a much greater chance. Conditional probability, and specifically **Bayes' rule**, is used for these kinds of problems.

Diagnosis from testing

Let's turn our attention to medicine. A growing practice in medicine is to claim that decision-making in medicine should be based on approaches that give us the best decisions. A search of PubMed texts for "evidence based medicine" found more than 91,944 (13 October 2021, an increase of thirteen thousand since last I looked (10 October, 2019). **Evidence based medicine** (EBM) is the "conscientious, explicit, judicious and reasonable use of modern, best evidence in making decisions about the care of individual patients" (Masic et al 2008). By **evidence**, we may mean results from quantitative, **systematic reviews**, **meta-analysis**, of research on a topic of medical interest, e.g., Cochrane Reviews.

Note:

Primary research refers to generating or collecting original data in pursuit of tests of hypotheses. Both systematic reviews and meta-analysis are **secondary research** or "research on research." As opposed to a literature review, systematic review make explicit how studies were searched for and included; if enough reasonably similar quantitative data are obtained through this process, the reviewer can combine the data and conduct an analysis to assess whether a treatment is effective (De Vries 2018).

As you know, no **diagnostic test** is 100% foolproof. For many reasons, test results come back positive when the person truly does not have the condition — this is a **false positive** result. Correctly identifying individuals who do not have the condition, which ideally means having a 0% false positive rate, is called the ***specificity*** of a test. Think of specificity in this way — provide the test 100 **true negative** samples (e.g., 100 samples from people who do not have cancer) — how many times out of 100 does the test correctly return a "negative"? If 99 times out of 100, then the specificity rate for this test is 99%, which is pretty good. But the test results mean more if the condition/disease is common; for rare conditions, even 99% is not good enough. Incorrect assignments are

rare, we believe, in part because the tests are generally quite accurate. However, what we don't consider is that detection and diagnosis from tests also depend on how frequent the incidence of the condition is in the population. Paradoxically, the lower the base rate, the poorer diagnostic value even a sensitive test may have.

To summarize our jargon for interpreting a test or assay, so far we have

- **True positive (a)**, the person has a disease and the test correctly identifies the person as having the disease.
- **False positive (b)**, test result incorrectly identifies disease; the person does not have the disease, but the test classifies the person as having the disease.
- **False negative (c)**, test result incorrectly classifies a person as not having disease, but the person actually has the disease.
- **True negative** (d), the person does not have the disease and the test correctly categorizes the person as not having the disease.
- **Sensitivity of test** is the proportion of persons who test positive and do have the disease (true positives): $Se = \frac{TP}{TP+FN}$
  If a test has 75% sensitivity, then out of 100 individuals who do have the disease, 75 will test positive (TP = true positive).
- **Specificity of a test** refers to the rate that a test correctly classifies a person that does not have the disease: $Sp = \frac{TN}{TN+FP}$
  If a test has 90% specificity, then out of 100 individuals who truly do not have the disease, 90 will test negative (TN = true negative).

A worked example. A 50-year-old male patient is in the doctor's office. The doctor is reviewing results from a diagnostic test, e.g., a FOBT — fecal occult blood test — a test used as a screening tool for colorectal cancer (CRC). The doctor knows that the test has a sensitivity of about 75% and specificity of about 90%. Prevalence of CRC in this age group is about 0.2%. Figure 7.3.4 shows our probability tree using our natural numbers approach.

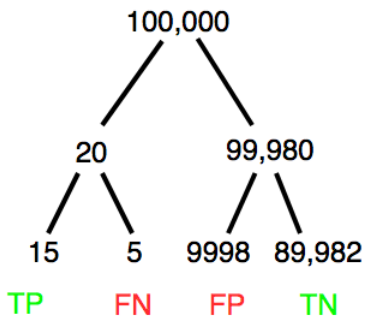


Figure 7.3.4: Probability tree for FOBT test. Desired test outcomes shown in green: TP stands for true positive and TN stands for true negative. Poor outcomes of a test shown in red: FN stands for false negative and FP stands for false positive.

The associated probabilities for the four possible outcomes of these kinds of tests (e.g., what's the probability of a person who has tested positive in screening tests actually having the disease?) are shown in Table 7.3.1.

Table 7.3.1. A 2 × 2 table of possible outcomes of a diagnostic test.

	Does person really have the disease?	
Test Result:	**Yes**	**No**
Positive	***a*** TP	***b*** FP
Negative	***c*** FN	***d*** TN

Bayes' rule is often given in probabilities,

$$p(D|\oplus) = \frac{p(D) \cdot p(\oplus|D)}{[p(D) \cdot p(\oplus) + p(ND) \cdot p(\oplus|ND)]}$$

=where truth is represented by either D (the person really does have the disease) or ND (the person really does not have the disease) and $\oplus$ is the symbol for "exclusive or" and reads "not D" in this example.

An easier way to see this is to use frequencies instead. Now, the formula is

$$p(disease|positive) = \frac{a}{a+b}$$

This is the simplified version of Bayes' rule, where a is the number of people who test positive and DO HAVE the disease and b is the number of people who test positive and DO NOT have the disease.

Standardized million

Where did the 100,000 come from? We discussed this in Chapter 2: it's a simple trick to adjust rates to the same population size. We use this to work with natural numbers instead of percentages or frequencies. You choose to correct to a standardized population based on the raw incidence rate. A rough rule of thumb:

Table 7.3.2. Relationship between standard population size and incidence rate.

Raw IR rate about	IR	Standard population
1/10	10%	1000
1/100	1%	10,000
1/1000	0.1%	100,000
1/10,000	0.01%	1,000,000
1/100,000	0.001%	10,000,000
1/1,000,000	0.0001%	100,000,000

The raw incident rate is simply the number of new cases divided by the total population.

Per capita rate

Yet another standard manipulation is to consider the average incidence per person, or per capita, rate. The Latin "per capita" translates to "by head" (Google translate), but in economics, epidemiology, and other fields it is used to reflect rates per person. Tuberculosis is a serious infectious disease of primarily the lungs. Incidence rates of tuberculosis in the United States have trended down since the mid-1900s: 52.6 per 100K in 1953 to 2.7 per 100K in 2019 (CDC). Corresponding per capita values are 5.26×10^{-4} and 2.7×10^{-5}, respectively. Divide the incidence rate by 100,000 to get the per-person rate.

Practice and introduce PPV and Youden's J

Let's break these problems down, and in doing so, introduce some terminology common to the field of "risk analysis" as it pertains to biology and epidemiology. Our first example considers the fecal occult blood test**,** FOBT, test. Blood in the stool may (or may not) indicate polys or colon cancer. (Park et al 2010).

Table 7.3.3. A 2 × 2 table of possible outcomes of FOBT test.

	Person really has the disease		
	Yes	**No**	
Positive	**15**	**9998**	$PPV = \frac{15}{(15+9998)} = 0.15\%$
Negative	**5**	**89,982**	$NPV = \frac{89982}{(89982+5)} = 99.99\%$

The table shown above will appear again and again throughout the course, but in different forms.

We want to know, how good is the test, particularly if the goal is early detection? This is conveyed by the **PPV**, **positive predictive value** of the test. Unfortunately, the prevalence of a condition is also at play: the lower the prevalence, the lower the PPV must be, because most positive tests will be false when population prevalence is low.

Youden (1950) proposed a now widely adopted index that summarizes how effective a test is. **Youden's J** is the sum of specificity and sensitivity minus one.

$$J = Se + Sp - 1$$

where Se stands for sensitivity of the test and Sp stands for sensitivity of the test.

Youden's J takes on values between 0, for a terrible test, and 1, for a perfect test. For our FOBT example, Youden's J was 0.65. This statistic looks like it's independent of prevalence, but its use as a **decision criterion** (e.g., a cutoff value, above which test is positive, below test is considered negative), assumes that the cost of misclassification (false positives, false negatives) are equal. Prevalence affects the number of false positives and false negatives for a given diagnostic test, so any decision criterion based on Youden's J will also be influenced by prevalence (Smits 2010).

Another worked example. A study on cholesterol-lowering drugs (statins) reported a relative risk reduction of death from cardiac event by 22%. This does not mean that for every 1000 people fitting the population studied, 220 people would be spared from having a heart attack. In the study, the death rate per 1000 people was 32 for the statin versus 41 for the **placebo** — recall that a placebo is a control treatment offered to overcome potential patient psychological bias (see Chapter 5.4). The absolute risk reduction due to statin is only $41 - 32$ or 9 in 1000 or 0.9%. By contrast, relative risk reduction is calculated as the ratio of the absolute risk reduction (9) divided by the proportion of patients who died without treatment (41), which is 22% (LIPID Study Group 1998).

Note that **risk reduction** is often conveyed as a **relative** rather than as an **absolute** number. The distinction is important for understanding arguments based in conditional probability. Thus, the question we want to ask about a test is summarized by **absolute risk reduction** (**ARR**) and **number needed to treat** (**NNT**), and for problems that include control subjects, **relative risk reduction** (**RRR**). We expand on these topics in the next section, 7.4 – Epidemiology: Relative risk and absolute risk, explained.

Evidence-Based Medicine

One culture change in medicine is the explicit intent to make decisions based on evidence (Masic et al 2008). Of course, the joke then is, well, what were doctors doing before, diagnosing without evidence? This comic strip xkcd offers one possible answer (Fig. 7.3.5).

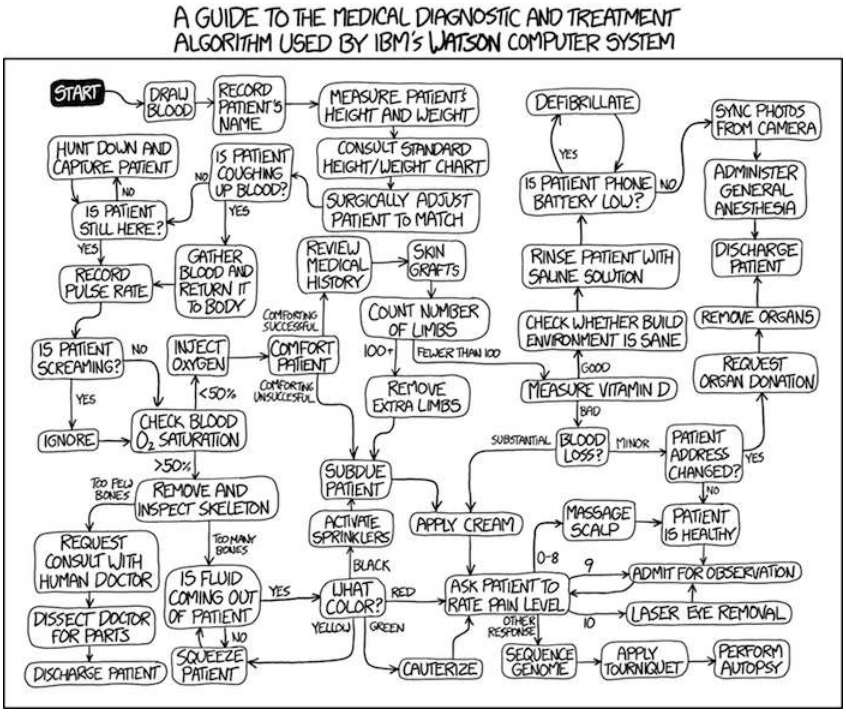


Figure 7.3.5: A summary of "evidence-based medical" decisions, perhaps? https://xkcd.com/1619/

As you can imagine, there's considerable reflection about the EBM movement (see discussions in response to Accad and Francis 2018, e.g., Goh 2018). More practically, our objective is for you to be able to work your way through word problems involving risk analysis. You can expect to be asked to calculate, or at least set up for calculation, any of the statistics listed above (e.g., false negative, false positive, etc.). Practice problems are listed at the end of this section, and additional problems are provided to you (Homework 4). You'll also want to check your work, and in any real analysis, you'd most likely want to use R.

Software

R has several epidemiology packages, and with some effort, can save you time. Another option is to run your problems in OpenEpi, a browser-based set of tools. OpenEpi is discussed with examples in the next section, 7.4.

Here, we illustrate some capabilities of the epiR package, expanded more also in the next section, 7.4. We'll use the example from Table 7.3.3.

R code:

```
library(epiR)
Table3 <- matrix(c(15, 5, 9998, 89982), nrow = 2, ncol = 2)
epi.tests(Table3)
```

R output:

```
         Outcome + Outcome -   Total
Test +          15      9998   10013
Test -           5     89982   89987
Total           20     99980  100000

Point estimates and 95% CIs:
-------------------------------------------------------------
Apparent prevalence *                       0.10 (0.10, 0.10)
True prevalence *                           0.00 (0.00, 0.00)
Sensitivity *                               0.75 (0.51, 0.91)
Specificity *                               0.90 (0.90, 0.90)
Positive predictive value *                 0.00 (0.00, 0.00)
Negative predictive value *                 1.00 (1.00, 1.00)
Positive likelihood ratio                   7.50 (5.82, 9.67)
Negative likelihood ratio                   0.28 (0.13, 0.59)
False T+ proportion for true D- *           0.10 (0.10, 0.10)
False T- proportion for true D+ *           0.25 (0.09, 0.49)
False T+ proportion for T+ *                1.00 (1.00, 1.00)
False T- proportion for T- *                0.00 (0.00, 0.00)
Correctly classified proportion *           0.90 (0.90, 0.90)
-------------------------------------------------------------
* Exact CIs
```

Oops! I wanted PPV, which by hand calculation was 0.15%, but R reported "0.00?" This is a **significant figure** reporting issue. The simplest solution is to submit `options(digits=6)` before the command, then save the output from `epi.tests()` to an object and use `summary()`. For example

```
options(digits=6)
myEpi <- epi.tests(Table3)
summary(myEpi)
```

And R returns

```
     statistic           est        lower        upper
1           ap  0.1001300000 0.0982761568  0.102007072
```

```
2           tp  0.0002000000 0.0001221693  0.000308867
3           se  0.7500000000 0.5089541283  0.913428531
4           sp  0.9000000000 0.8981238085  0.901852950
5      diag.ac  0.8999700000 0.8980937508  0.901823014
6      diag.or 27.0000000000 9.8110071871 74.304297826
7         nndx  1.5384615385 1.2265702376  2.456532054
8       youden  0.6500000000 0.4070779368  0.815281481
9       pv.pos  0.0014980525 0.0008386834  0.002469608
10      pv.neg  0.9999444364 0.9998703379  0.999981958
11      lr.pos  7.5000000000 5.8193604069  9.666010707
12      lr.neg  0.2777777778 0.1300251423  0.593427490
13      p.rout  0.8998700000 0.8979929278  0.901723843
14       p.rin  0.1001300000 0.0982761568  0.102007072
15      p.tpdn  0.1000000000 0.0981470498  0.101876192
16      p.tndp  0.2500000000 0.0865714691  0.491045872
17      p.dntp  0.9985019475 0.9975303919  0.999161317
18      p.dptn  0.0000555636 0.0000180416  0.000129662
```

There we go — `pv.pos` reported as 0.0014980525, which, after turning to a percent and rounding, we have 0.15%. Note also the additional statistics provided — a good rule of thumb — always try to save the output to an object, then view the object, e.g., with `summary()`. Refer to help pages for additional details of the output (`?epi.tests`).

What about R Commander menus?

Note: Fall 2023

I have not been able to run the EBM plugin successfully! It simply returns an error message — on data sets which have in the past performed perfectly. Thus, until further notice, do not use the EBM plugin. Instead, use commands in the epiR package. I'm leaving the text here on the chance the error with the plugin is fixed.

Rcmdr has a plugin that will calculate ARR, RRR and NNT. The plugin is called `RcmdrPlugin.EBM` (Leucuta et al 2014) and it would be downloaded as for any other package via R.

Download the package from your selected R mirror site, then start R Commander.

```
install.packages("RcmdrPlugin.EBM")
```

From within R Commander (Fig. 7.3.6), select

Tools → Load Rcmdr plug-in(s)…

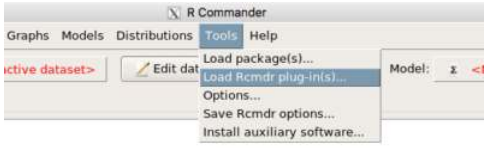


Figure 7.3.6: To install an Rcmdr plugin, first go to Rcmdr → Tools → Load Rcmdr plug-in(s)…

Next, select from the list the plug-in you want to load into memory, in this case, `RcmdrPlugin.EBM` (Fig. 7.3.7).

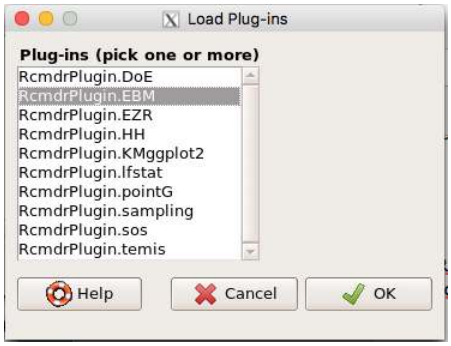


Figure 7.3.7: Select the Rcmdr plugin, then click the “OK” button to proceed.

Restart Rcmdr again (Fig. 7.3.8),

Figure 7.3.8: Select “Yes” to restart R Commander and finish installation of the plug-in.

and the menu “EBM” should be visible in the menu bar (Fig. 7.3.9).

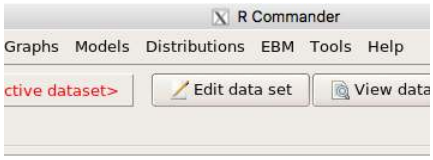


Figure 7.3.9: Copy and Paste Caption here. (Copyright; author via source)

Note that you will need to repeat these steps each time you wish to work with a plug-in, unless you modify your `.RProfile` file. See

Rcmdr → Tools → Save Rcmdr options…

Clicking on the EBM menu item brings up the template for the Evidence Based Medicine module. We’ll mostly work with **2 × 2 tables** (e.g., see Table 7.3.1) , so select the “Enter two-way table…” option to proceed (Fig. 7.3.10).

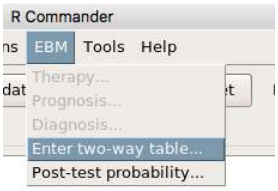


Figure 7.3.10: Select “Enter two-way table…”.

And finally, Figure 7.3.11 shows the two-way table entry cells along with options. We’ll try a problem by hand, then use the EBM plugin to confirm and gain additional insight.

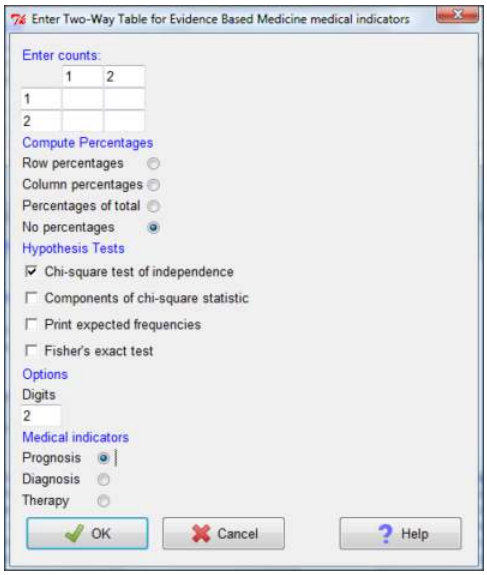


Figure 7.3.11: Two-way table Rcmdr EBM plug-in.

For assessing how good a test or assay is, use the Diagnosis option in the EBM plugin. For situations with treated and control groups, use Therapy option. For situations in which you are comparing exposure groups (e.g., smokers vs non-smokers), use the Prognosis option.

Example

Here's a simple one (problem from Gigerenzer 2002).

About 0.01% of men in Germany with no known risk factors are currently infected with HIV. If a man from this population actually has the disease, there is a 99.9% chance the tests will be positive. If a man from this population is not infected, there is a 99.9% chance that the test will be negative. *What is the chance that a man who tests positive actually has the disease?*

Start with the reference, or base population (Figure 7.3.12). It's easy to determine the rate of HIV infection in the population if you use numbers. For 10,000 men in this group, exactly one man is likely to have HIV (0.0001 × 10,000), whereas 9,999 would not be infected.

For the man who has the disease it's virtually certain that his results will be positive for the virus (because the sensitivity rate = 99.9%). For the other 9,999 men, one will test positive (the false positive rate = 1 – specificity rate = 0.01%).

Thus, for this population of men, for every two who test positive, one has the disease and one does not, so the probability even given a positive test is only 100 × 1/2 = 50%. This would also be the test's Positive Predictive Value.

Note that if the base rate changes, then the final answer changes!

It also helps to draw a tree to help you determine the numbers (Fig. 7.3.12)

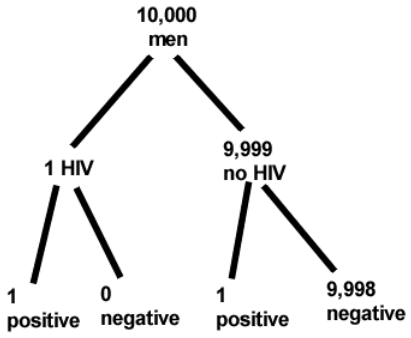


Figure 7.3.12: Draw a probability tree to help with the frequencies.

From our probability tree in Figure 7.3.12 it is straightforward to collect the information we need.

- Given this population, how many are expected to have HIV? **Two**.
- Given the specificity and sensitivity of the assay for HIV, how many persons from this population will test positive? **Two**.
- For every positive test result, how many men from this population will actually have HIV? **One**.

Thus, given this population with the known risk associated, the probability that a man testing positive actually has HIV is 50% $\left(=\frac{1}{(1+1)}\right)$.

Use the EBM plugin. Select two-way table, then enter the values as shown in Fig. 7.3.13

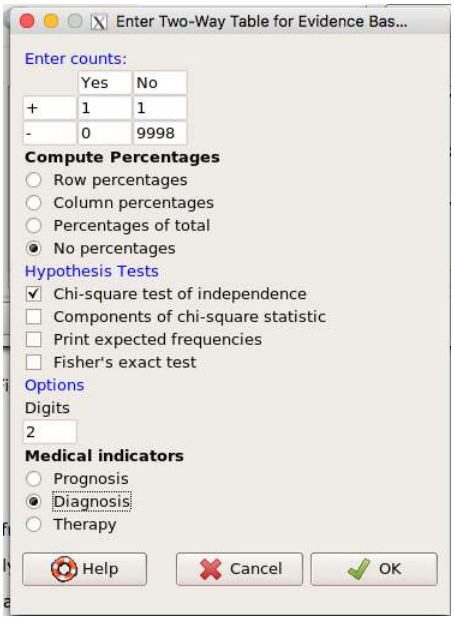


Figure 7.3.13: EBM plugin with data entry.

Select the "Diagnosis" option — we are answering the question: How probable is a positive result given information about sensitivity and specificity of a diagnosis test? The results from the EBM functions are given below.

```
Rcmdr> .Table
      Yes No
   +  1    1
   -  0 9998
```

```
Rcmdr> fncEBMCrossTab(.table=.Table, .x='', .y='', .ylab='', .xlab='',
Rcmdr+ .percents='none', .chisq='1', .expected='0', .chisqComp='0', .fisher='0',
Rcmdr+ .indicators='dg', .decimals=2)
```

```
# Notations for calculations
        Disease + Disease -
Test +      "a"         "b"
Test -      "c"         "d"
```

```
# Sensitivity (Se) = 100 (95% CI 2.5 - 100) %. Computed using formula: a / (a + c)
# Specificity (Sp) = 99.99 (95% CI 99.94 - 100) %. Computed using formula: d / (b + d
# Diagnostic accuracy (% of all correct results) = 99.99 (95% CI 99.94 - 100) %. Compu
# Youden's index = 1 (95% CI 0.02 - 1). Computed using formula: Se + Sp - 1
# Likelihood ratio of a positive test = 9999 (95% CI 1408.63 - 70976.66). Computed us
```

```
# Likelihood ratio of a negative test = 0 (95% CI 0 - NaN). Computed using formula: (
# Positive predictive value = 50 (95% CI 1.26 - 98.74) %. Computed using formula: a /
# Negative predictive value = 100 (95% CI 99.96 - 100) %. Computed using formula: d /
# Number needed to diagnose = 1 (95% CI 1 - 40.91). Computed using formula: 1 / [Se -
```

Note that the formulas used to calculate Sensitivity, Specificity, etc., follow our Table 7.3.1 (compare to "Notations for calculations"). The use of EBM provides calculations of our confidence intervals.

Questions

1. The sensitivity of the fecal occult blood test (FOBT) is reported to be 0.68. What is the False Negative Rate?
2. The specificity of the fecal occult blood test (FOBT) is reported to be 0.98. What is the False Positive Rate?
3. For men between 50 and 54 years of age, the rate of colon cancer is 61 per 100,000. If the false negative rate of the fecal occult blood test (FOBT) is 10%, how many persons who have colon cancer will test negative?
4. For men between 50 and 54 years of age, the rate of colon cancer is 61 per 100,000. If the false positive rate of the fecal occult blood test (FOBT) is 10%, how many persons who do not have colon cancer will test positive?
5. A study was conducted to see if mammograms reduced mortality (data from Table 5-1 p. 60 Gigerenzer (2002)). What is the RRR?

Mammogram	Deaths/1000 women
No	4
Yes	3

6. A study was conducted to see if mammograms reduced mortality (data from Table 5-1 p. 60 Gigerenzer (2002)). What is the NNT?

Mammogram	Deaths/1000 women
No	4
Yes	3

7. Does supplemental Vitamin C decrease risk of stroke in Type II diabetic women? In a study conducted on 1923 women, a total of 57 women had a stroke. Of the 57, 14 were in the normal Vitamin C level and 32 were in the high Vitamin C level. What is the NNT between normal and high supplemental Vitamin C groups?
8. Sensitivity of a test is defined as
   A. False Positive Rate
   B. True Positive Rate
   C. False Negative Rate
   D. True Negative Rate
9. Specificity of a test is defined as
   A. False Positive Rate
   B. True Positive Rate
   C. False Negative Rate
   D. True Negative Rate
10. In thinking about the results of a test of a null hypothesis, Type I error rate is equivalent to
    A. False Positive Rate
    B. True Positive Rate
    C. False Negative Rate
    D. True Negative Rate
11. During the Covid-19 pandemic, number of reported cases each day were published. For example, 155 cases were reported for 9 October 2020 by Department of Health. What is the raw incident rate?

7.4: Epidemiology relative risk and absolute risk, explained

Introduction

Epidemiology is the study of patterns of health and illness of populations. An important task in an epidemiology study is to identify risks associated with disease. Epidemiology is a crucial discipline used to inform about possible effective treatment approaches, health policy, and about the etiology of disease.

Please review terms presented in section 7.1 before proceeding. RR and AR are appropriate for **cohort-control** and **cross-sectional** studies (see 2.4 and 5.4) where base rates of exposure and unexposed or numbers of affected and non-affected individuals (**prevalence**) are available. Calculations of **relative risk** (RR) and **relative risk reduction** (RRR) are specific to the sampled groups under study whereas **absolute risk** (AR) and **absolute risk reduction** (ARR) pertain to the reference population. Relative risks are specific to the study, absolute risks are generalized to the population. **Number needed to treat** (NNT) is a way to communicate absolute risk reductions.

An example of ARR and RRR risk calculations using natural numbers

Clinical trials are perhaps the essential research approach (Sibbald and Roland 1998; Sylvester et al 2017); they are often characterized with a binary outcome. Subjects either get better or they do not. There are many ways to represent risk of a particular outcome, but where possible, using **natural numbers** is generally preferred as a means of communication. Consider the following example (pp 34-35, Gigerenzer 2015): What is the benefit of taking a cholesterol-lowering drug, Pravastatin, on the risk of deaths by heart attacks and other causes of mortality? Press releases (e.g., Maugh 1995), from the study stated the following:

> "… the drug pravastatin reduced … deaths from all causes 22%".

A subsequent report (Skolbekken 1998) presented the following numbers (Table 7.4.1).

Table 7.4.1. Reduction in total mortality (5 year study) for people who took Pravastatin compared to those who took placebo.

		Deaths per 1000 people with high cholesterol (> 240 mg/dL)	No deaths	Cumulative incidence
Treatment	Pravastatin (n = 3302)	a = 32	b = 3270	$CI_e e$
	Placebo (n = 3293)	c = 41	d = 3252	CI_u

where **cumulative incidence** refers to the number of new events or cases of disease divided by the total number of individuals in the population at risk.

Do the calculations of risk

The risk reduction (RR), or the number of people who die without treatment (placebo) minus those who die with treatment (Pravastatin), $41 - 32 = 9$.

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = 0.91$$

The **cumulative incidence** in the exposed (treated) group, CI_e, is $\frac{32}{32+3270} = 0.0097$, and cumulative incidence in the unexposed (control) group, CI_u, is $\frac{41}{41+3252} = 0.01245$. We can calculate another statistic called the **risk ratio**,

$$RR = \frac{CI_e}{CI_u} = 0.78$$

Because the risk ratio is less than one, we interpret that statins reduce the risk of mortality from heart attack. In other words, statins lowered the risk by 0.78.

But is this risk reduction meaningful?

Now, consider the absolute risk reduction (ARR) is $0.9\% = 100\% \times \frac{9}{1000}$.

Relative risk reduction, or the absolute risk reduction divided by the proportion of patients who die without treatment, is $22\% = 100\% \times 9 \div 41$.

Conclusion: high cholesterol may contribute to increased risk of mortality, but the rate is very low in the population as a whole (the ARR).

Another useful way to communicate benefit is to calculate the Number Needed to Treat (NNT), or the number of people who must receive the treatment to save (benefit) one person. The ideal NNT is a value of one (1), which would be interpreted as everyone improves who receives the treatment. By definition, NNT must be positive; however, a resulting negative NNT would suggest the treatment may cause harm, i.e., number needed to harm (NNH).

For this example, the NNT is

$$\frac{1}{\frac{9}{1000}} = 111$$

Therefore, to benefit one person, 111 need to be treated. The flip side of the implications of NNT is that although one person may benefit by taking the treatment, $111 - 1 = 110$ will take the treatment and will NOT RECEIVE THE BENEFIT, but do potentially get any side effect of the treatment.

Confidence interval for NNT is derived from the Confidence interval for ARR

For a sample of 100 people drawn at random from a population (which may number in the millions), if we then repeat the NNT calculation for a different sample of 100 people, do we expect the first and second NNT estimates to be exactly the same number? No, but we do expect them to be close, and we can define what we mean by close as we expect each estimate to be within certain limits. While we expect the second calculation to be close to the first estimate, we would be surprised if it was exactly the same. And so, which is the correct estimate, the first or the second? They both are, in the sense that they both estimate the parameter NNT (a property of a population).

We use **confidence intervals** to communicate where we believe the true estimate for NNT to be. Confidence Intervals (CI) allow us to assign a probability to how certain we are about the statistic and whether it is likely to be close to the true value (Altman 1998, Bender 2001). We will calculate the 95% CI for the ARR using the **Wald method**, then take the inverse of these estimates for our 95% CI. The Wald method assumes normality.

For CI of ARR, we need sample size for control and treatment groups; like all confidence intervals, we need to calculate the standard error of the statistic, in this, case, the standard error (SE) for ARR is approximately

$$SE_{(p_1 - p_2)} = \sqrt{\frac{p_1\left(1 - p_1\right)}{n_1} + \frac{p_2\left(1 - p_2\right)}{n_2}}$$

where SE is the standard error for ARR. For our example, we have

$$SE_{(p_1 - p_2)} = \sqrt{\frac{0.041\left(1 - 0.041\right)}{1000} + \frac{0.032\left(1 - 0.032\right)}{1000}}$$

The 95% CI for ARR is approximately $ARR \pm 2 \times SE_{(p_1 - p_2)}$.

For the Wald estimate, replace the 2 with $z = 1.965$, which comes from the normal table for z at $\frac{0.95}{2}$. Why the 2 in the equation? Because it is plus or minus so we divide the frequency 0.95 in half) and for our example, we have $0.009 \pm 2 \times SE_{(p_1 - p_2)} = (-0.0078, 0.0258)$ and the inverse for NNT CI is $(-128, 38)$.

Our example exemplifies the limitation of the Wald approach (cf. Altman 1998): our confidence interval includes zero, and doesn't even include our best estimate of NNT (111).

Note:

By now you should see differences for results by direct input of the numbers into R and what you get by the natural numbers approach. In part this is because we round in our natural number calculations — remember, while it makes more sense to communicate about whole numbers (people) and not fractions (fractions of people!), rounding through the calculations adds

error to the final value. As long as you know the difference and the relevance between approximate and exact solutions, this shouldn't cause concern.

Software: epiR

R has many epidemiology packages, `epiR` and `epitools` are two. Most of the code presented stems from epiR.

We need to know about our study design in order to tell the functions which statistics are appropriate to estimate. For our statin example, the design was prospective cohort (i.e., `cohort.count` in epiR package language), not case-control or cross-sectional (review in Chapter 5.4).

```
library(epiR)
Table1 <- matrix(c(32,3270,41,3252), 2, 2, byrow=TRUE, dimnames = list(c("Statin", "P.
Table1
        Died Lived
Statin    32  3270
Placebo   41  3252
epi.2by2(Table1, method="cohort.count", outcome = "as.columns")
```

R output:

```
           Outcome + Outcome - Total           Inc risk *
Exposed +         32      3270  3302 0.97  (0.66 to 1.37)
Exposed -         41      3252  3293 1.25  (0.89 to 1.69)
Total             73      6522  6595 1.11  (0.87 to 1.39)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                          0.78 (0.49, 1.23)
Inc odds ratio                          0.78 (0.49, 1.24)
Attrib risk in the exposed *           -0.28 (-0.78, 0.23)
Attrib fraction in the exposed (%)    -28.48 (-103.47, 18.88)
Attrib risk in the population *        -0.14 (-0.59, 0.32)
Attrib fraction in the population (%) -12.48 (-37.60, 8.05)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 1.147 Pr>chi2 = 0.284
Fisher exact test that OR = 1: Pr>chi2 = 0.292
Wald confidence limits
CI: confidence interval
* Outcomes per 100 population units
```

The risk ratio we calculated by hand is shown in green in the R output, along with other useful statistics (see \?epi2x2 for help with these additional terms) not defined in our presentation.

We explain results of chi-square goodness of fit (Ch 9.1) and Fisher exact (Ch 9.5) tests in Chapter 9. Suffice to say here, we interpret the p-value (Pr) = 0.284 and 0.292 to indicate that there is no association between mortality from heart attacks with or without the statin (i.e., the Odds Ratio, OR, not statistically different from one).

Wait! Where's NNT and other results?

Use another command in epiR package, `epi.tests()` , to determine the specificity, sensitivity, and positive (or negative) predictive value.

```
epi.tests(Table1)
```

R returns:

```
          Outcome + Outcome - Total
Test +           32      3270  3302
Test -           41      3252  3293
Total            73      6522  6595

Point estimates and 95% CIs:
-------------------------------------------------------------
Apparent prevalence *                    0.50 (0.49, 0.51)
True prevalence *                        0.01 (0.01, 0.01)
Sensitivity *                            0.44 (0.32, 0.56)
Specificity *                            0.50 (0.49, 0.51)
Positive predictive value *              0.01 (0.01, 0.01)
Negative predictive value *              0.99 (0.98, 0.99)
Positive likelihood ratio                0.87 (0.67, 1.13)
Negative likelihood ratio                1.13 (0.92, 1.38)
False T+ proportion for true D- *        0.50 (0.49, 0.51)
False T- proportion for true D+ *        0.56 (0.44, 0.68)
False T+ proportion for T+ *             0.99 (0.99, 0.99)
False T- proportion for T- *             0.01 (0.01, 0.02)
Correctly classified proportion *        0.50 (0.49, 0.51)
-------------------------------------------------------------
* Exact CIs
```

Additional statistics are available by saving the output from `epi2x2()` or `epitests()` to an object, then using `summary()`. For example save output from `epi.2by2(Table1, method="cohort.count", outcome = "as.columns")` to object `myEpi`, then

```
summary(myEpi)
```

look for NNT in the R output

```
$massoc.detail$NNT.strata.wald
       est    lower    upper
1 -362.377 -128.038  436.481
```

Thus, the NNT was 362 (compared to the 111 we got by hand) with a 95% Confidence interval between -436 and $+128$ (make it positive because it is a treatment improvement.)

Note:

Strata (L. layers) refer to subgroups, for example, sex or age categories. Our examples are not presented as subgroup analysis, but epiR reports by name strata.

`epiR` reports a lot of additional statistics in the output and for clarity, I have not defined each one, just the basic terms we need for BI311. As always, see help pages (e.g., `\?epi.2x2` or `\?epitests`)for more information about structure of an R

command and the output.

We're good, but we can work the output to make it more useful to us.

Improve output from epiR

For starters, if we set `interpret=TRUE` instead of the default, `interpret=FALSE`, `epiR` will return a richer response.

```
fit <- epi.2by2(dat = as.table(Table1), method = "cohort.count", conf.level = 0.95, u
fit
```

R output. In addition to the table of coefficients (above), `interpret=TRUE` provides more context, shown below:

```
Measures of association strength:
The outcome incidence risk among the exposed was 0.78 (95% CI 0.49 to 1.23) times les

The outcome incidence odds among the exposed was 0.78 (95% CI 0.49 to 1.24) times les

Measures of effect in the exposed:
Exposure changed the outcome incidence risk in the exposed by -0.28 (95% CI -0.78 to

Number needed to treat for benefit (NNTB) and harm (NNTH):
The number needed to treat for one subject to be harmed (NNTH) is 362 (NNTH 128 to in

Measures of effect in the population:
Exposure changed the outcome incidence risk in the population by -0.14 (95% CI -0.59
```

That's quite a bit. Another trick is to get at the table of results. We install a package called `broom`, which includes a number of ways to handle output from R functions, including those in the epiR package. Broom takes from the TidyVerse environment; tables are stored as tibbles.

```
library(broom)

# Test statistics
tidy(fit, parameters = "stat")
```

R output:

```
# A tibble: 3 × 4
term statistic df p.value
<chr> <dbl> <dbl> <dbl>
1 chi2.strata.uncor 1.15 1 0.284
2 chi2.strata.yates 0.909 1 0.340
3 chi2.strata.fisher NA NA 0.292
```

We can convert the tibbles into our familiar data.frame format, and then select only the statistics we want.

```
# Measures of association
fitD <- as.data.frame(tidy(fit, parameters = "moa")); fitD
```

R output shows all 15 measures of association!

```
                  term     estimate     conf.low    conf.high
1       RR.strata.wald    0.7783605    0.4914679   1.23272564
2     RR.strata.taylor    0.7783605    0.4914679   1.23272564
3      RR.strata.score    0.8742994    0.6584540   1.10340173
4       OR.strata.wald    0.7761915    0.4876209   1.23553616
5     OR.strata.cfield    0.7761915           NA           NA
6      OR.strata.score    0.7761915    0.4894450   1.23093168
7        OR.strata.mle    0.7762234    0.4718655   1.26668220
8    ARisk.strata.wald   -0.2759557   -0.7810162   0.22910484
9   ARisk.strata.score   -0.2759557   -0.8000574   0.23482532
10     NNT.strata.wald -362.3770579 -128.0383246 436.48140194
11    NNT.strata.score -362.3770579 -124.9910314 425.84844829
12  AFRisk.strata.wald   -0.2847517   -1.0347210   0.18878949
13  PARisk.strata.wald   -0.1381661   -0.5933541   0.31702189
14  PARisk.strata.piri   -0.1381661   -0.3910629   0.11473067
15 PAFRisk.strata.wald   -0.1248227   -0.3760279   0.08052298
```

We can call out just the statistics we want from this table by calling to the specific elements in the data.frame (rows, columns).

```
fitD[c(1,4,7,9,12),]
```

R output:

```
                  term   estimate   conf.low  conf.high
1       RR.strata.wald  0.7783605  0.4914679  1.2327256
4       OR.strata.wald  0.7761915  0.4876209  1.2355362
7        OR.strata.mle  0.7762234  0.4718655  1.2666822
9   ARisk.strata.score -0.2759557 -0.8000574  0.2348253
12 AFRisk.strata.wald -0.2847517 -1.0347210  0.1887895
```

Software: epitools

Another useful R package for epidemiology is `epitools`, but it comes with its own idiosyncrasies. We have introduced the standard 2 × 2 format, with a, b, c, and d cells defined as in Table 7.4.1 above. However, `epitools` does it differently, and we need to update the matrix. By default, `epitools` has the unexposed group (control) in the first row and the non-outcome (no disease) is in the first column. To match our **a,b,c,** and **d** matrix, use the `epitools` command to change this arrangement with the `rev()` argument. Now, the analysis will use the contingency table on the right where the exposed group (treatment) is in the first row and the outcome (disease) is in the first column (h/t M. Bounthavong 2021). Once that's accomplished, `epitools` returns what you would expect.

Calculate relative risk:

```
risk1 <- 32 / (3270 + 32)
risk2 <- 41 / (3525 + 41)
risk1 - risk2
```

and R returns:

```
-0.00180638
```

Calculate the odds ratio:

```
library(epitools)
oddsratio.wald(Table1, rev = c("both"))
```

and R returns:

```
$data
              Outcome
Predictor Disease2 Disease1 Total
 Exposed2      517       36   553
 Exposed1      518       11   529
Total         1035       47  1082

$measure
           odds ratio with 95% C.I.
Predictor   estimate      lower      upper
 Exposed2  1.0000000         NA         NA
 Exposed1  0.3049657 0.1535563  0.6056675

$p.value
two-sided
Predictor  midp.exact  fisher.exact    chi.square
Exposed2           NA            NA            NA
Exposed1 0.0002954494  0.0003001641  0.0003517007
```

Odds ratio is highlighted in green.

Software: OpenEpi

R is fully capable of delivering the calculations you need, but sometimes you just want a quick answer. Online, the OpenEpi tools at https://www.openepi.com/ can be used for homework problems. For example, working with count data in 2×2 format, select Counts > 2×2 table from the side menu to bring up the data form (Fig. 7.4.1).

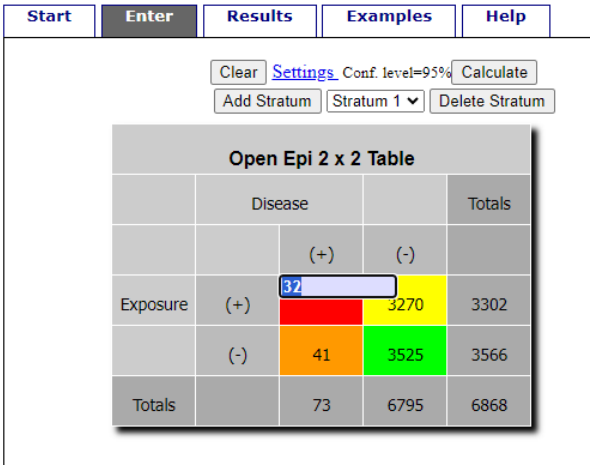
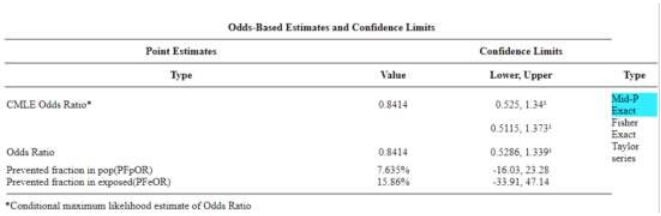


Figure 7.4.1: Data entry for 2×2 table at openepi.com.

Once the data are entered, click on the **Calculate** button to return a suite of results.



Odds-Based Estimates and Confidence Limits

Point Estimates		Confidence Limits	
Type	Value	Lower, Upper	Type
CMLE Odds Ratio*	0.8414	0.525, 1.34[1]	Mid-P Exact
		0.5115, 1.373[1]	Fisher Exact
Odds Ratio	0.8414	0.5286, 1.339[1]	Taylor series
Prevented fraction in pop(PFpOR)	7.635%	-16.03, 23.28	
Prevented fraction in exposed(PFeOR)	15.86%	-33.91, 47.14	

*Conditional maximum likelihood estimate of Odds Ratio

Figure 7.4.1: Results for 2×2 table at openepi.com.

Software: RcmdrPlugin.EBM

Note: Fall 2023

I have not been able to run the EBM plugin successfully! It simply returns an error message — on data sets which have in the past performed perfectly. Thus, until further notice, do not use the EBM plugin. Instead, use commands in the `epiR` package.

This isn't the place nor can I be the author to discuss what evidence based medicine (EBM) entails (cf. Masic et al. 2008), or what its shortcomings may be (Djulbegovic and Guyatt 2017). Rcmdr has a nice plugin, based on the `epiR` package, that will calculate ARR, RRR and NNT as well as other statistics. The plugin is called `RcmdrPlugin.EBM`

```
install.packages("RcmdrPlugin.EBM", dependencies=TRUE)
```

After acquiring the package, proceed to install the plug-in. Restart Rcmdr, then select Tools and Rcmdr Plugins (Fig 7.4.3).

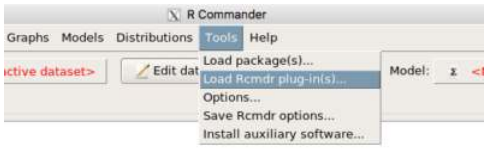


Figure 7.4.3: Rcmdr: Tools → Load Rcmdr plugins…

Find the EBM plug-in, then proceed to load the package (Fig. $PageIndex4$).

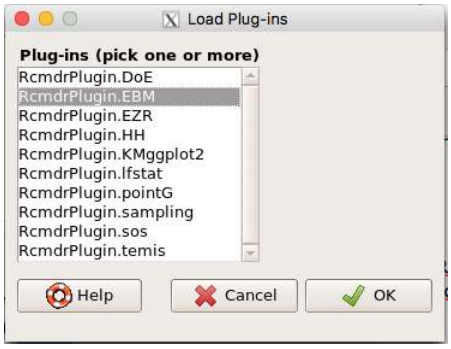


Figure 7.4.4: Rcmdr plug-ins available (after first downloading the files from an R mirror site).

Restart Rcmdr again and the menu "EBM" should be visible in the menu bar. We're going to enter some data, so choose the Enter two-way table… option in the EBM plug-in (Fig 5)

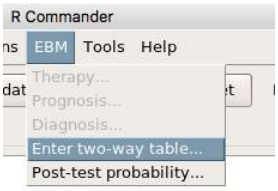


Figure 7.4.5: R Commander EBM plug-in, enter 2×2 table menus

To review, we have the following problem, illustrated with natural numbers and probability tree (Fig. 7.4.6).

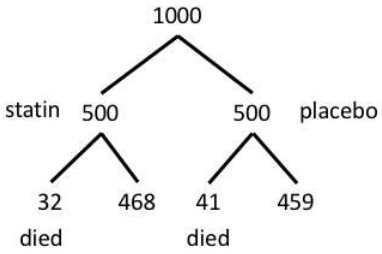


Figure 7.4.6: Illustration of probability tree for the statin problem.

Now, let's enter the data into the EBM plugin. For the data above I entered the counts as

	Lived	Died
Statin	468	32
Placebo	459	41

and selected the "Therapy" medical indicator (Fig. 7.4.7)

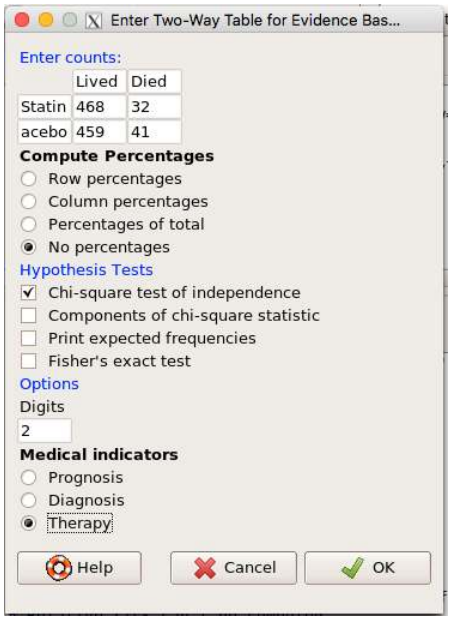


Figure 7.4.7: EBM plugin with two-way table completed for the statin problem.

The output from EBM plugin was as follows. I've added index numbers in brackets so that we can point to the output that is relevant for our worked example here.

```
(1) .Table <- matrix(c(468,32,459,41), 2, 2, byrow=TRUE, dimnames = list(c('Drug', 'P.
(2) fncEBMCrossTab(.table=.Table, .x='', .y='', .ylab='', .xlab='', .percents='none',
```

R output begins by repeating the commands used, here marked by lines (1) and (2). The statistics we want follow in the next several lines of output.

```
(3) Pearson's Chi-squared test data: .Table X-squared = 1.197, df = 1, p-value = 0.27
(4) # Notations for calculations Event + Event -Treatment "a" "b" Control "c" "d"
(5)# Absolute risk reduction (ARR) = -1.8 (95% CI -5.02 - 1.42) %. Computed using for
(6)# Relative risk = 1.02 (95% CI 0.98 - 1.06) %. Computed using formula: [c / (c + d
(7)# Odds ratio = 1.31 (95% CI 0.81 - 2.11). Computed using formula: (a / b) / (c / d
(8) # Number needed to treat = -55.56 (95% CI 70.29 - Inf). Computed using formula: 1
9)# Relative risk reduction = -1.96 (95% CI -5.57 - 1.53) %. Computed using formula:
(10)# To find more about the results, and about how confidence intervals were compute
```

In summary, we found no difference between statin and placebo (P-value $= 0.2739$), and an ARR of -1.8%.

Questions

Data from a case-control study on alcohol use and esophageal cancer (Tuyns et al (1977), example from Gerstman 2014). Cases were men diagnosed with esophageal cancer from a region in France. Controls were selected at random from electoral lists from the same geographical region. Use this data for questions 1–4.

Table 7.4.2. Data from case-control study on alcohol use and esophageal cancer.

	Esophageal Cancer		
Alcohol grams/day	Cases	Noncases	Total
> 80	96	109	205
< 80	104	666	770
Total	200	775	975

1. What was the null hypothesis? Be able to write the hypothesis in symbolic form and as a single sentence.
2. What was the alternate hypothesis? Be able to write the hypothesis in symbolic form and as a single sentence.
3. What was the observed frequency of subjects with esophageal cancer in this study? And the observed frequency of subjects without esophageal cancer?
4. Estimate Relative Risk, Absolute Risk, NNT, and Odds ratio.
    1. Which is more appropriate, RR or OR? Justify your decision.
5. The American College of Obstetricians and Gynecologists recommends that women with an average risk of breast cancer (BC) over 40 get an annual mammogram. Nationally, the sensitivity of mammography is about 68% and specificity of mammography is about 75%. Moreover, mammography involves exposure of women to radiation, which is known to cause mutations. Given that the prevalence of BC in women between 40 and 49 is about 0.1%, please evaluate the value of this recommendation by completing your analysis.
   A) In this age group, how many women are expected to develop BC?
   B) How many False negative would we expect?
   C) How many positive mammograms are likely to be true positives?
6. “Less than 5% of women with screen-detectable cancers have their lives saved,” (quote from BMC Med Inform Decis Mak. 2009 Apr 2;9:18. doi: 10.1186/1472-6947-9-18): Using the information from question 5, what is the Number Needed to Treat for mammography screening?

7.5: Odds ratio

Introduction

We introduced the concept of odds 7.1: Epidemiology definitions. As a reminder, odds are a way to communicate the chance (likelihood) that a particular event will take place. Odds are calculated as the number of individuals with the event divided by the number of individuals without the event.

Odds ratio is a measure of **effect size** for the association between two binary (yes/no) variables. It is the ratio of the odds of an event occurring in one group to the odds of the same event happening in another group. The odds ratio (OR) is a way to quantify the strength of association between one condition and another.

> **Note:**
>
> Effect size — the size of the difference between groups — is discussed further in Chapter 9.2 and Chapter 11.4.

How are odds ratios calculated? The probabilities are conditional; recall finding the conditional probability of some event ***A***, given the occurrence of some other event ***B***.

Let p_{y-y} equal probability of the event occurring (y = Yes) in ***A***, p_{y-n} equal probability of the event not occurring (n = No) in ***A***, p_{n-y} equal probability of the event occurring in ***B***, and p_{n-n} equal probability of the event not occurring in ***B***.

		A	
		Yes	No
B	Yes	p_{y-y}	p_{y-n}
	No	p_{n-y}	p_{n-n}

These sum to one: $p_{y-y} + p_{y-n} + p_{n-y} + p_{n-n} = 1$

The conditional probabilities are

		A	
		Yes	No
B	Yes	$\frac{p_{y-y}}{p_{y-y} + p_{n-n}}$	$\frac{p_{y-n}}{p_{y-y} + p_{y-n}}$
	No	$\frac{p_{n-y}}{p_{n-y} + p_{n-n}}$	$\frac{p_{n-n}}{p_{n-y} + p_{n-n}}$

and finally then, the **odds ratio** (OR) is

$$OR = \frac{p_{y-y} \cdot p_{n-n}}{p_{y-n} \cdot p_{n-y}}$$

If you have the raw numbers you can calculate the odds ratio directly, too.

		A	
		Yes	No
B	Yes	a	b
	No	c	d

and the odds ratio is then

$$OR = \frac{a \div b}{c \div d}$$

or, equivalently,

$$OR = \frac{a \cdot d}{b \cdot c}$$

Example

Comparing proportions is a frequent need in court. Gray (2002) provided an example from Title IX of the Education Act of 1972 case *Cohen v. Brown University*. Under the Act, discrimination based on gender is prohibited. The case concerned participation in collegiate athletics by women. The case data were that of the 5722 undergraduate students, 51% were women, but of the 987 athletes, only 38% were women. A mosaic plot shows graphically these proportions (Fig. 7.5.1, males in red bars, females in yellow bars).

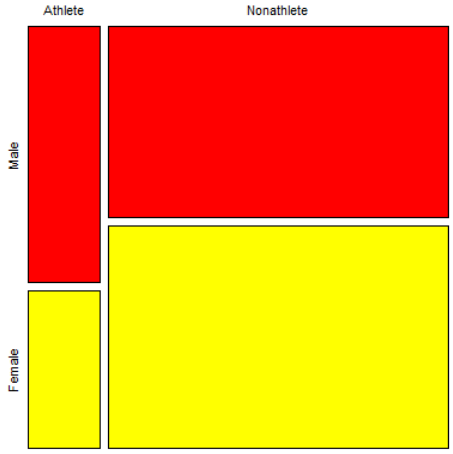


Figure 7.5.1: Mosaic plot of athletes to non-athletes in college. Males red, females yellow, data from Gray 2002.

Alternatively, use a **Venn diagram** to describe the distribution (Fig. 7.5.2). Circles that overlap show regions of commonality.

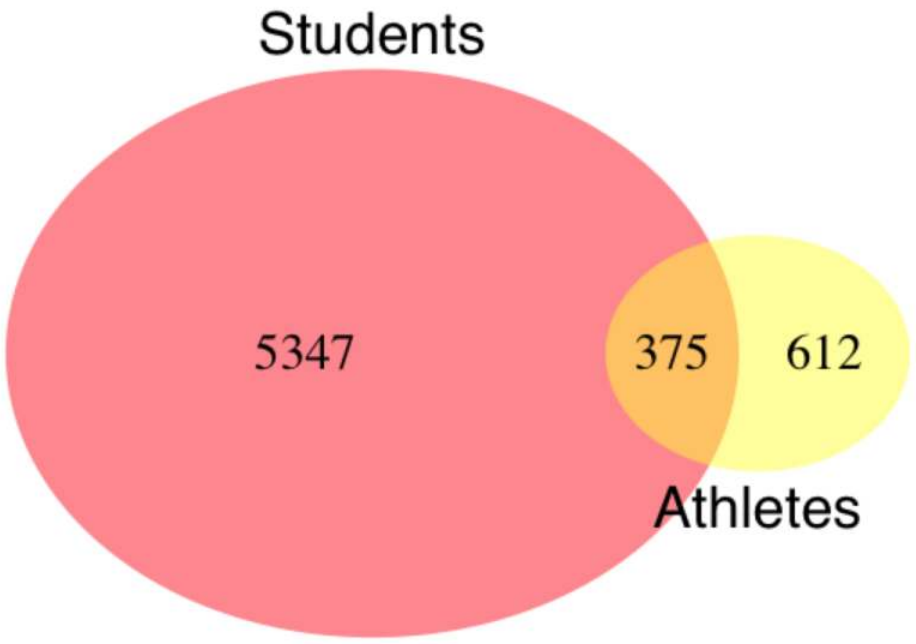


Figure 7.5.2: Venn Diagram of athletes to non-athletes in Brown University. Female athletes $(n = 375)$, male athletes $(n = 612)$, data from Gray 2002.

where the orange region represents $\text{Students} \cap \text{Female Athletes}$.

R code for the Venn diagram was

```
library(VennDiagram)
area1 = 5722
area2 = 987
cross.area = 375
draw.pairwise.venn(area1,area2,cross.area,category=c("Students","Athletes"),
euler.d = TRUE, scaled = TRUE, inverted = FALSE, print.mode = "percent",
fill=c("Red","Yellow"),cex = 1.5, lty="blank", cat.fontfamily = rep("sans", 2),
cat.cex = 1.7, cat.pos = c(0, 180), ext.pos=0)
```

The question raised before the court was whether these proportions meet the demand of "substantially proportionate." What exactly the law means by "substantially proportionate" was left to the courts and the lawyers to work out (Gray 2002). Title IX suggests that "substantially proportionate" is a statistical problem and the two sides of the argument must address the question from that perspective.

What is the chance that an undergraduate student was an athlete and female? 38%. And the chance that an undergraduate student was an athlete and male? 62%. Clearly 38% is not 62%; did the plaintiffs have a case?

Graphs like Figure 7.5.1 and Figure 7.5.2 help communicate but can't provide a sense of whether the differences are important. Let's start by looking at the numbers. Working with the proportions, we have the following breakdown for numbers of students (Table 7.5.1) or as proportions (Table 7.5.2).

Table 7.5.1. Gray's raw data displayed in a 2×2 format.

		Athletes	
		Yes	*No*
Undergraduates	*Male*	612	2192
	Female	375	2543

Together, the numbers total 5,722.

The Odds Ratio (OR) would be

$$OR = \frac{612 \cdot 2543}{2192 \cdot 375} = 1.89$$

Or from the proportions (Table 7.5.2):

Table 7.5.2. Data from Table 7.5.1 as proportions.

		Athletes	
		Yes	*No*
Undergraduates	*Male*	0.107	0.383
	Female	0.066	0.444

Adding all of these frequencies together equals 1. Carry out the calculation of odds (Table 7.5.3), which shows the conditional probabilities in bold.

Table 7.5.3. Odds calculated from Table 7.5.2 inputs.

		Athletes	
		Yes	*No*

Undergraduates	*Male*	$\mathbf{0.218} = \frac{0.107}{0.107+0.383}$	$\mathbf{0.782} = \frac{0.383}{0.107+0.383}$
	Female	$\mathbf{0.129} = \frac{0.066}{0.066+0.444}$	$\mathbf{0.871} = \frac{0.444}{0.066+0.444}$

Calculate the odds ratio:

$$OR = \frac{0.2182 \cdot 0.871}{0.129 \cdot 0.871} = 1.89$$

Thankfully, whether we use the raw number format or the proportion format, we got the same results!

Interpretation. Because the Odds Ratio (OR) is greater than 1, males students were more likely to be athletes than female students. If there was no difference in proportion of male and female athletes, the odds ratio would be close to one. That is a test of statistical inference (e.g., a contingency table), but for now, if one is included in the confidence interval, then this would be evidence that there was no difference between the proportions.

Relative risk v. odds ratio

We introduced another way to quantify this association as the Relative Risk (RR) and Absolute Risk Reductions in the previous section. Both can be used to describe the risk of the treatment (exposed) group relative to the control (nonexposed) group. RR is the ratio of the treated to control group. OR is the ratio between odds of treated (exposed) and control (nonexposed). What's the difference? OR is more general — it can be used in situations in which the researcher chooses the number of affected individuals in the groups and, therefore, the base rate or prevalence of the condition in the population is not known or is not represented in the group makeup, whereas RR is appropriate when prevalence is known (this is a general point, but see Schechtman 2002 for a nice discussion).

The odds ratio is related to relative risk, but not over the entire range of possible risk. Odds of an event is simply the number of individuals with the event divided by the number without the event. Odds of an event therefore can range from zero (event cannot occur) to infinity (event must occur). For example, odds of eight (1.89:1) means that nearly two male students were student athletes at Brown University for every one female student.

In contrast, the risk of an event occurring is the number of individuals with the event divided by the total number of people at risk of having that event. Risk is expressed as a percentage (Davies et al 1998). Thus, for our example, odds of 1.89:1 correspond to a risk of 1.89 divided by (1 + 1.89), which equals 65%.

To get the relative risk we can use

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

For our example, this comes out to 1.7%.

In this example we could use either odds or relative risk; the key distinction is that we knew how many events happened in both groups. If this information is missing for one group (e.g., control group of the case-control design), then only the odds ratio would be appropriate.

From cumulative wisdom in the literature (e.g., Tamhane et al 2107), if prevalence is less than ten percent, $OR \approx RR$. We can relate RR and OR as

$$RR = OR \cdot \frac{1 + \frac{n_{2,1}}{n_{2,2}}}{1 + \frac{n_{1,1}}{n_{1,2}}}$$

where $n_{1,1}$ and $n_{2,1}$ are the frequency with the condition for group 1 and group 2, respectively, and $n_{1,2}$ and $n_{2,2}$ are the frequency without the condition for group 1 and group 2, respectively. For the examples on this page, group 1 is the treatment group and group 2 is the control group.

Hazard ratio

The hazard ratio is the ratio of hazard rates. **Hazard rates** are like the relative risk rates, but are specific to a period of time. Hazard rates come from a technique called **Survival Analysis** (introduced in Chapter 20.9). Survival analysis can be thought of as following a group of subjects over time until something (the event) happens. By following two groups, perhaps one group exposed to a suspected carcinogen vs. another group matched in other respects except the exposure, at the end of the trial, we'll have two hazard rates: the rate for the exposed group and the rate for the control group. If there is no difference, then the hazard ratio will be one.

Hazard ratios are more appropriate for clinical trials; relative risk is more appropriate for observational studies.

For a hazard ratio, it is often easier to think of it as a probability (between 0 to 1). To translate a hazard ratio to a probability, use the following equation:

$$p = \frac{hazardratio}{1 + hazardratio}$$

Questions

1. Distinguish between odds ratio, relative risk, and hazard ratio.
2. Refer to problem 4 introduced in 7.4 – Epidemiology: Relative risk and absolute risk, explained.

7.6: Confidence intervals

Introduction

Although I've already presented the concept (e.g., Chapter 3.4), and equations for confidence intervals of risk estimates (Chapter 7.4, Chapter 7.5), we'll expand on the idea of **confidence intervals**. Confidence intervals are a central part of meeting one of the main objectives of statistics, that is, **estimation**. We will review how to calculate confidence intervals for proportions and for NNT. These intervals are available in `epiR` package and automatically returned in `RcmdrPlugin.EBM` .

There are three components of statistical analysis:

1. Estimation
2. Inference
3. Modeling

Inference refers to statisftical hypothesis testing; we ask questions of observations — do men (Rice et al 1999) and women (Fisher et al 2012) differ for blood glucose levels following a bout of aerobic exercise? T-tests, analysis of variance (ANOVA), chi-square, correlation, regression are types of statistical procedures used to do statistical inference. **Modeling**, on the other hand, refers to procedures used to relate cause and effect. Many of the statistical procedures one uses for inference are also used to build statistical models (ANOVA, regression). Studies may intend to either test some hypothesis (inference) or to provide a predictive equation (modeling). But most studies that relate observations gathered from an experiment are obliged to also report statistics, and this is the realm of **estimation**. Estimates of the mean and standard deviation, for example, would be typical statistics one expects to find in a report. We call these descriptive statistics, and together with graphics, descriptive statistics are the chief way we describe our results.

Confidence interval for proportions

A proportion is the fraction of individuals in a population with some characteristic. The characteristic might be HIV positive, for example. This would be called the population proportion and it would be a parameter of interest. In reality, we calculate a sample proportion and therefore estimate the population proportion with error. We can calculate the confidence interval (CI) of the proportion to communicate the precision of our estimate. For proportions, we use the binomial distribution — either a sample has the characteristic of interest or it does not; there are only two possibilities. There are a variety of ways to go here, and the simplest is to use a normal approximation. This will work well provided the sample size was reasonably large and the proportion is not close to zero or one, that is, we invoke the Central Limit Theorem here. Although the outcomes are binomial, the error is assumed to be normally distributed. The **Wald confidence interval** for p is

$$95\%\ CI = \hat{p} \pm z\sqrt{\frac{1}{n} \cdot \hat{p}\,(1-\hat{p})}$$

where $\hat{p}$ is the proportion of individuals with the characteristic (also called successes), z is the percentile from the normal distribution that corresponds to $1 - \frac{1}{2}\alpha$. For 95% CI, then $\alpha = 0.05$, which would mean $z = 1.965$. (See standard normal table.) Of course, if making the normal approximation for the binomial is not appropriate, the CI is less than ideal. The binomial, after all, is a discrete distribution whereas the normal distribution is continuous, so errors will enter, particularly for low sample numbers.

Other approaches may be used to get better estimates of CI for proportions, including **Wilson score intervals** and **Jeffrey Intervals** (Agresti and Coull 1998). See R package `propCIs` .

Because a statistic like the mean or a calculation of absolute or relative risk reduction are calculated from samples drawn from a population, the estimate comes with **error**. The error is basically this – if we calculate a statistic like number needed to treat (NNT) or its converse, the number needed to harm (NNH), we need to communicate to the reader how **precise** our estimate is. Estimation has to do with accuracy, error, and precision.

Confidence interval for ARR

The ARR is simply $\frac{c}{n_2} - \frac{a}{n_1}$,

where n_1 is the number of treated or exposed individuals for which the event occurred and n_2 is the number of untreated or unexposed individuals which the event occurred.

	Event happened	**Event did not happen**

Treated or Exposed	a	b
Control or Not exposed	c	d

Our data from the Brown University example in Chapter 7.5 were $a = 612$, $b = 2192$, $c = 375$, and $d = 2543$. \[SE_{ARR} = \sqrt{\left(\frac{a}{n_{1}} \cdot \frac{1 - \frac{a}{n_1}}{n_{1}}\right) + \left(\frac{c}{n_{2} \cdot \frac{1 - \frac{c}{n_{2}}}{n_{2}}\right) } \nonumber\]

and the 95% confidence interval is then approximately $ARR \pm 2 \cdot SE_{ARR}$.

The "2" is only approximate; you need to use $z = 1.965$, the z-value at a probability value of 0.9725 (which comes from the Normal Table).

Confidence interval for NNT

If we calculate the NNT for a sample of 100 people drawn at random from a population (which may number in the millions), then repeat the NNT calculation for a different sample of 100 people, do we expect the first and second NNT estimates to be EXACTLY the same number? No, but we do expect them to be close, and we can define what we mean by close as we expect each estimate to be within certain limits. While we expect the second calculation to be close to the first estimate, we would be surprised if it was EXACTLY the same. And so, which is the correct estimate, the first or the second? They both are, in the sense that they both estimate the parameter NNT (a property of a population). But we can do better than two estimates. Confidence Intervals (CI) allow us to assign a probability to how certain we are about the statistic and whether it is likely to be close to the true value. We will

For CI of NNT, we need sample size for control and treatment groups; like all confidence intervals, we need to calculate the standard error of the statistic: in this case, the standard error (SE) for NNT.

```
SE = sqrt(risk placebo * (1 - risk placebo) / (# in placebo group) + risk treatment *
```

where SE is the standard error for NNT

The CI is approximately then $NNT \pm 2 \times SE$.

Note that the "2" is only approximate; you need to use $z = 1.965$, the z-value at a probability value of 0.9725 (which comes from the Standard Normal Table).

Odds ratio Standard error and 95% confidence interval

Like any statistic we can calculate, an estimate of odds ratio should be accompanied by the confidence limit. The standard error may be calculated with the following formula:

$$SE\{\ln(OR)\} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

R code:

```
seOdds <- sqrt(sum(1/612, 1/2192,1/375,1/2543))
```

In this equation, ln refers to the natural logarithm. An estimate for the 95% confidence interval is

$$lower\ limit = \exp(\ln(OR)) - 1.96 \cdot SE\{\ln(OR)\}$$

where exp is the exponential function e^x. In our example the lower limit was 1.64.

R code:

```
exp(log(1.89,base=exp(1)) - 1.96*seOdds)
```

For the upper limit, we calculate

$$upper\ limit = \exp(\ln(OR)) + 1.96 \cdot SE\{\ln(OR)\}$$

In our example the upper limit was 2.19.

R code:

```
exp(log(1.89,base=exp(1)) + 1.96*seOdds)
```

Thus, our estimate was 1.89 and the 95% confidence interval was $(1.64, 2.19)$ which does not include one. Therefore, we conclude that the male and female student groups are statistically different.

Questions

1. Instead of 95% confidence interval, obtain the 99% confidence interval for an odds ratio of 1.89.

2. What would be the value of Z used for a 99% confidence interval for ARR and NNT?

7.7: Chapter 7 References and Suggested Readings

Accad, M., & Francis, D. (2018). Does evidence based medicine adversely affect clinical judgment?. *BMJ: British Medical Journal (Online)*, *362*.

Agresti, A., Coull, B. A. (1998). Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52:119-126.

Altman, D. G. (1998). Confidence intervals for the number needed to treat. *BMJ*, 317(7168), 1309-1312.

Bender, R. (2001). Calculating confidence intervals for the number needed to treat. *Controlled clinical trials*, 22(2), 102-110.

Berry, D. A. (2008). The science of doping. *Nature* 454(7205):692-3.

Bewick, V., Cheek, L., Ball, J. (2003). Statistics review 8: Qualitative data – tests of association. *Critical Care*, 8:46-53.

Carling, C. L. L., Kristoffersen, D. T., Montori, V. M., Herrin, J., Schünemann, H. J., Treweek, S., Akl, E. A., Oxman, A. D. (2009). The Effect of Alternative Summary Statistics for Communicating Risk Reduction on Decisions about Taking Statins: A Randomized Trial. *PLoS Medicine*, 6(8): e1000134

Davies, H. T., Crombie, I. K., Tavakoli, M. (1998). When can odds ratios mislead? BMJ, 316: 989-991

Decker, R. C. (2008). *A brief review of the basic principles of epidemiology, Chapter 2 in Field Epidemiology*, 3rd edition, Gregg MB editor. Oxford University Press

de Vrieze, Jop. (2018) The metawars. *Science* 361: 1184-1188.

Fisher, G., Hunter, G. R., Gower, B. A. (2012). Aerobic exercise training conserves insulin sensitivity for 1 yr following weight loss in overweight women. *J. Appl. Physiol.*, 112, 688-693.

Galinsky, A. M., Zelaya, C. E., Barnes, P. M., & Simile, C. (2022). Selected health conditions among native Hawaiian and Pacific Islander adults: United States, 2014. *Cancer*, *23*, 2.

Gerstman, B. B. (2014). *Basic biostatistics: Statistics for public health practice*, 2nd ed. Jones & Bartlett Learning.

Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You*. Simon & Schuster.

Gray, M. W. (2002). Cramming for court: teaching statistics to litigators. ICOTS6: The Sixth International Conference on Teaching Statistics.

Hastings, D/. (2003). The Challenger Disaster. https://ocw.mit.edu/courses/engineering-systems-division/esd-10-introduction-to-technology-and-policy-fall-2006/readings/challenger.pdf.

Johnson, K. M. (2017). Using Bayes' rule in diagnostic testing: a graphical explanation. *Diagnosis*, *4*(3), 159-167.

Krebs, J. R. (2014). Risk, uncertainty and regulation. Philosophical Transactions. Series a, Mathematical, Physical, and Engineering Sciences. 369: 4842-52. PMID 22042900 DOI: 10.1098/rsta.2011.0174

Leucuta, D. C., Călinici, T., Drugan, T., Istrate, D., & Achimas, A. (2014). Graphical User Interface Extension in R Commander for Evidence Based Medicine Indicators. *Applied Medical Informatics.*, *35*(3), 11-16.

LIPID Study Group (1998). Prevention of Cardiovascular Events and Death with Pravastatin in Patients with Coronary Heart Disease and a Broad Range of Initial Cholesterol Levels. *NEJM*, 339:1349-1357.

Masic, I., Miokovic, M., & Muhamedagic, B. (2008). Evidence based medicine–new approaches and challenges. *Acta Informatica Medica*, 16(4), 219.

Maugh, T.H., II. (1995, November 16). Anti-Cholesterol Drug Cuts Heart Attacks, Study Shows : Health: Pravastatin reduced risks for healthy men, research finds. Experts predict new era in prevention. Los Angeles Times, Retrieved from https://www.latimes.com/archives/la-...794-story.html.

Park, D. I., Ryu, S., Kim, Y. H., Lee, S. H., Lee, C. K., Eun, C. S., & Han, D. S. (2010). Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at average risk undergoing colorectal cancer screening. *American Journal of Gastroenterology*, 105(9), 2017-2025.

Parikh, R., Parikh, S., Arun, E., Thomas, R. (2009). Likelihood ratios: Clinical application in day-to-day practice. *Indian Journal of Opthamology* 57(3): 217–221.

Rice, B., Janssen, I., Hudson, R., Ross, R. (1999). Effects of aerobic or resistance exercise and/or diet on glucose tolerance and plasma insulin levels in obese men. *Diabetes Care* 22:684-691

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use?. *Value in health*, *5*(5), 431-436.

Sibbald, B., & Roland, M. (1998). Understanding controlled trials. Why are randomised controlled trials important? *BMJ* 316(7126), 201.

Skolbekken, J. A. (1998). Communicating the risk reduction achieved by cholesterol reducing drugs. *BMJ*, 316(7149), 1956-1958.

Smits, N. (2010). A note on Youden's J and its cost ratio. *BMC medical research methodology*, *10*(1), 1-4.

Sylvester, R. J., Canfield, S. E., Lam, T. B., Marconi, L., MacLennan, S., Yuan, Y., … & Hernandez, V. (2017). Conflict of evidence: resolving discrepancies when findings from randomized controlled trials and meta-analyses disagree. *European urology*, 71(5), 811-819.

Tamhane, A. R., Westfall, A. O., Burkholder, G. A., & Cutter, G. R. (2016). Prevalence odds ratio versus prevalence ratio: choice comes with consequences. *Statistics in medicine*, *35*(30), 5730-5735.

US Department of Health and Human Services. (2013). *Principles of Epidemiology in Public Health Practice, Third Edition: An Introduction to Applied Epidemiology and Biostatistics*. Atlanta, Georgia, USA. Available on the website: http://www.cdc.gov/ophss/csels/dsepd/SS1978.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32-35.

CHAPTER OVERVIEW

8: Inferential Statistics

Introduction

Statistical methods are important in biology because results of experiments are usually not clear-cut and therefore tests to support decisions between competing hypotheses are needed.

We will limit ourselves to a general discussion with examples, but beginning in this chapter, we start our introductions of specific types of statistical tests. As a reminder, our statistical philosophy is **frequentist** and follows the **Null Hypothesis Significant Testing** or NHST approach. Discussion of **Bayesian statistical** approaches are included as appropriate.

Thus, all statistical tests we will talk about share the following requirements or properties.

1. The type of data we have dictates which test or tests are appropriate.
2. We start with a clear description of the **null hypotheses**.
3. Set the **Type I error rate**, **alpha** (α). By convention, 5% is often used (Cowles and David 1982)
4. We must be aware of the assumptions our statistical tests make and what, if any, modifications to them we can make.
5. Correct computation of the **test statistic** and **degrees of freedom**.
6. Comparison of the **critical value** and the test statistic value, with interpretation and significance testing (**p-value**, **Bayesian Factor**, cf. discussion in Goodman 2008).

We can provide a flow-chart of these steps (Fig. 8.1).

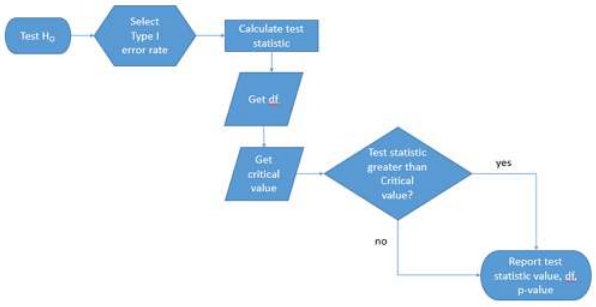


Figure 8.1: NHST decision flow chart.

While we want to avoid the impression that statistical analysis is simply a matter of following a step-by-step protocol as in Fig. 8.1, it nevertheless may be helpful to think of it as such, understanding all the while that there are caveats and assumptions that accompany the choices we make while following the protocol.

8.1: The null and alternative hypotheses

Introduction

Classical statistical parametric tests — **t-tests** (one sample t-test, independent sample-t-test), **analysis of variance** (ANOVA), **correlation**, and **linear regression**— and nonparametric tests like χ^2 **(chi-square: goodness of fit and contingency table),** share several features that we need to understand. It's natural to see all the details as if they are specific to each test, but there's a theme that binds all of the classical statistical inference in order to make claim of "statistical significance."

- a calculated **test statistic**
- **degrees of freedom** associated with the calculation of the test statistic
- a **probability value** or **p-value** which is associated with the test statistic, assuming a **null hypothesis** is "true" in the population from which we sample.
  - Note that as discussed in (Chapter 8.2), this is not strictly the interpretation of p-value, but a shorthand for how likely the data is to fit the null hypothesis. P-value alone can't tell us about "truth."
- in the event we reject the null hypothesis, we provisionally accept the **alternative hypothesis**.

Statistical Inference in the NHST Framework

By inference, we mean to imply some formal process by which a conclusion is reached from data analysis of outcomes of an experiment. The process at its best leads to conclusions based on evidence. In statistics, evidence comes about from the careful and reasoned application of statistical procedures and the evaluation of probability (Abelson 1995).

Formally, statistics is rich in inference process. We begin by defining the **classical frequentist,** aka Neyman-Pearson approach, to inference, which involves the pairing of two kinds of statistical hypotheses: the null hypothesis (H_O) and the alternate hypothesis (H_A). Whether we accept the hull hypothesis or not is evaluated against a decision criterion, a fixed **statistical significance level** (Lehmann 1992). Significance level refers to the setting of a **p-value threshold** before testing is done. The threshold is often set to Type I error of 5% (Cowles & Davis 1982), but researchers should always consider whether this threshold is appropriate for their work (Benjamin et al 2017).

This inference process is referred to as Null Hypothesis Significance Testing, NHST. Additionally, a probability value will be obtained for the test outcome or test statistic value. In the Fisherian **likelihood** tradition, the magnitude of this statistic value can be associated with a probability value, the p-value, of how likely the result is given that the null hypothesis is "true". (Again, keep in mind that this is not strictly the interpretation of p-value, it's a shorthand for how likely the data is to fit the null hypothesis. P-value alone can't tell us about "truth", per our discussion in Chapter 8.2.)

Note:

About **-logP**. P-values are traditionally reported as a decimal, like 0.000134, in the **closed (set) interval** (0,1) — p-values can never be exactly zero or one. The smaller the value, the less the chance our data agree with the null prediction. Small numbers like this can be confusing, particularly if many p-values are reported, like in many genomics works, e.g., GWAS studies. Instead of reporting vanishingly small p-values, studies may report the **negative $\log_{10}$ p-value**, or **-logP**. Instead of small decimal numbers, large numbers are reported; the larger, the more chance our data is against the null hypothesis. Thus, our p-value becomes 3.87 -logP.

R code

```
-1*log(0.000134,10)
[1] 3.872895
```

Why $\log_{10}$ and not some other base transform? Just that $\log_{10}$ is convenient — powers of 10.

The **antilog** of 3.87 returns our p-value:

```
> 10^(-1*3.872895)
[1] 0.0001340001
```

For convenience, here is a partial p-value -logP transform table.

P-value	-logP
0.1	1
0.01	2
0.001	3
0.0001	4

On your own, complete the table for -logP values of 5 through 10. See Question 7 below.

NHST Workflow

We presented in the introduction to Chapter 8 without discussion a simple flow chart to illustrate the process of decision. Here, we repeat the flow chart diagram and follow with descriptions of the elements.

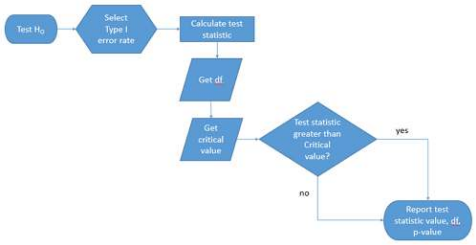


Figure 8.1.1: Flow chart of inductive statistical reasoning.

What's missing from the flow chart is the very necessary caveat that interpretation of the null hypothesis is associated with two kinds of error, Type I error and Type II error. These points and others are discussed in the following sections.

We start with the hypothesis statements. For illustration we discuss hypotheses in terms of comparisons involving just two groups, also called **two-sample tests**. **One-sample tests**, in contrast, refer to scenarios where you compare a sample statistic to a population value. Extending these concepts to more than two samples is straightforward, but we leave that discussion to Chapters 12 – 18.

Null hypothesis

By far the most common application of the null hypothesis testing paradigm involves the comparisons of different treatment groups on some outcome variable. These kinds of null hypotheses are the subject of Chapters 8 through 12.

The **Null hypothesis** (H_O) is a statement about the comparisons, e.g., between a sample statistic and the population, or between two treatment groups. The former is referred to as a **one-tailed test** whereas the latter is called a **two-tailed test**. The null hypothesis is typically "no statistical difference" between the comparisons.

For example, a one-sample, two-tailed null hypothesis.

$$H_O : \bar{X} = \mu$$

and we read it as "there is no statistical difference between our sample mean and the population mean." For the more likely case in which no population mean is available, we provide another example, a two-sample, two-tailed null hypothesis:

$$H_O : \bar{X}_1 = \bar{X}_2$$

Here, we read the statement as "there is no difference between our two sample means." Equivalently, we interpret the statement as both sample means estimate the same population mean.

$$H_O : \bar{X}_1 = \bar{X}_2 = \mu$$

Under the **Neyman-Pearson approach** to inference we have two hypotheses: the null hypothesis and the alternate hypothesis. The null hypothesis was defined above.

> **Note:**
>
> Tails of a test are discussed further in chapter 8.4.

Alternative hypothesis

Alternative hypothesis (H_A): If we conclude that the null hypothesis is false, or rather and more precisely, we find that *we provisionally fail to reject the null hypothesis*, then we *provisionally accept the alternative hypothesis*. The view then is that something other than random chance has influenced the sample observations. Note that the pairing of null and alternative hypotheses covers all possible outcomes. We do not, however, say that we have evidence for the alternative hypothesis under this statistical regimen (Abelson 1995). We tested the null hypothesis, not the alternative hypothesis. Thus, it is incorrect to write that, having found a statistical difference between two drug treatments, say aspirin and acetaminophen for relief of migraine symptoms, it is not correct to conclude that we have proven the case that acetaminophen improves improves symptoms of migraine sufferers.

For the one-sample, two-tailed null hypothesis, the alternative hypothesis is

$$H_A : \bar{X} \neq \mu$$

and we read it as "there is a statistical difference between our sample mean and the population mean." For the two-sample, two-tailed null hypothesis, the alternative hypothesis would be

$$H_A : \bar{X}_1 \neq \bar{X}_2$$

and we read it as "there is a statistical difference between our two sample means."

Alternative hypothesis often may be the research hypothesis

It may be helpful to distinguish between technical hypotheses, scientific hypothesis, or the equality of different kinds of treatments. Tests of technical hypotheses include the testing of statistical assumptions like **normality assumption** (see Chapter 13.3) and **homogeneity of variances** (Chapter 13.4). The results of inferences about technical hypotheses are used by the statistician to justify selection of parametric statistical tests (Chapter 13). The testing of some scientific hypothesis like whether or not there is a positive link between lifespan and insulin-like growth factor levels in humans (Fontana et al 2008), like the link between lifespan and IGFs in other organisms (Holtzenberger et al 2003), can be further advanced by considering multiple hypotheses and a test of nested hypotheses and evaluated either in Bayesian or likelihood approaches (Chapter 16 and Chapter 17).

How to interpret the results of a statistical test

Any number of statistical tests may be used to calculate the value of the **test statistic**. For example, a one-sample t-test may be used to evaluate the difference between the sample mean and the population mean (Chapter 8.5) or the independent sample t-test may be used to evaluate the difference between means of the control group and the treatment group (Chapter 10). The test statistic is the particular value of the outcome of our evaluation of the hypothesis and it is associated with the p-value. In other words, given the assumption of a particular probability distribution, in this case the t-distribution, we can associate a probability, the p-value, that we observed the particular value of the test statistic and the null hypothesis is true in the reference population.

By convention, we determine **statistical significance** (Cox 1982; Whitley & Ball 2002) by assigning ahead of time a decision probability called the **Type I error rate**, often given the symbol α (alpha). The practice is to look up the **critical value** that corresponds to the outcome of the test with degrees of freedom like your experiment and at the Type I error rate that you selected. The **Degrees of Freedom** (DF, df, or sometimes noted by the symbol v), are the number of independent pieces of information available to you. Knowing the degrees of freedom is a crucial piece of information for making the correct tests. Each statistical test has a specific formula for obtaining the independent information available for the statistical test. We first were introduced to DF when we calculated the sample variance with the **Bessel correction**, $n-1$, instead of dividing through by n. With df in hand, the value of the test statistic is compared to the critical value for our null hypothesis. If the test statistic is smaller than the critical value, we fail to reject the null hypothesis. If, however, the test statistic is greater than the critical value, then we provisionally reject the null hypothesis. This critical value comes from a probability distribution appropriate for the kind of sampling and properties of the measurement we are using. In other words, the rejection criterion for the null hypothesis is set to a critical value, which corresponds to a known probability, the Type I error rate.

Before proceeding with yet another interpretation, and hopefully a less technical discussion about test statistics and critical values, we need to discuss the two types of statistical errors. The Type I error rate is the statistical error assigned to the probability that we may reject a null hypothesis as a result of our evaluation of our data when in fact in the reference population, the null hypothesis is, in fact, true. In Biology we generally use Type I error $\alpha = 0.05$ level of significance. We say that the probability of obtaining the observed value AND H_O is true is 1 in 20 (5%) if $\alpha = 0.05$. Put another way, we are willing to reject the Null Hypothesis when there is only a 5% chance that the observations could occur and the Null hypothesis is still true. Our test statistic is associated with the p-value; the critical value is associated with the Type I error rate. If and only if the test statistic value equals the critical value will the p-value equal the Type I error rate.

The second error type associated with hypothesis testing is β, the **Type II statistical error rate**. This is the case where we accept or fail to reject a null hypothesis based on our data, but in the reference population, the situation is that indeed, the null hypothesis is actually false.

Thus, we end with a concept that may take you a while to come to terms with — there are four, not two possible outcomes of an experiment.

Outcomes of an experiment

What are the possible outcomes of a comparative experiment\? We have two treatments: one in which subjects are given a treatment and the other, in which subjects receive a placebo. Subjects are followed and an outcome is measured. We calculate the descriptive statistics aka summary statistics, means, standard deviations, and perhaps other statistics, and then ask whether there is a difference between the statistics for the groups. So, two possible outcomes of the experiment, correct\? If the treatment has no effect, then we would expect the two groups to have roughly the same values for means, etc., in other words, any difference between the groups is due to chance fluctuations in the measurements and not because of any systematic effect due to the treatment received. Conversely, then if there is a difference due to the treatment, we expect to see a large enough difference in the statistics so that we would notice the systematic effect due to the treatment.

Actually, there are four, not two, possible outcomes of an experiment, just as there were four and not two conclusions about the results of a clinical assay. The four possible outcomes of a test of a statistical null hypothesis are illustrated in Table 8.1.1.

Table 8.1.1. When conducting hypothesis testing, four outcomes are possible.

		H_O in the population	
		True	False
Result of statistical test	Reject H_O	Type I error with probability equal to α (alpha)	Correct decision, with probability equal to $1-\beta$ (1 – beta)
	Fail to reject the H_O	Correct decision with probability equal to $1-\alpha$ (1 – alpha)	Type II error with probability equal to β (beta)

In the actual population, a thing happens or it doesn't. The null hypothesis is either true or it is not. But we don't have access to the reference population, we don't have a census. In other words, there is truth, but we don't have access to the truth. We can weight, assigned as a probability or p-value, our decisions by how likely our results are given the assumption that the truth is indeed "no difference."

If you recall, we've seen a table like Table 8.1.1 before in our discussion of conditional probability and risk analysis (Chapter 7.3). We made the point that statistical inference and the interpretation of clinical tests are similar (Browner and Newman 1987). From the perspective of ordering a **diagnostic test**, the proper null hypothesis would be that the patient does not have the disease. For your review, here's that table (Table 8.1.2).

Table 8.1.2. Interpretations of results of a diagnostic or clinical test.

		Does the person have the disease?	
		Yes	No
Result of the diagnostic test	Positive	Sensitivity of the test (**a**)	False positive (**b**)
	Negative	False negative (**c**)	Specificity of the test (**d**)

Thus, a positive diagnostic test result is interpreted as rejecting the null hypothesis. If the person actually does not have the disease, then the positive diagnostic test is a false positive.

Questions

1. Match the corresponding entries in the two tables. For example, which outcome from the inference/hypothesis table matches *specificity of the test*?
2. Find three sources on the web for definitions of the p-value. Write out these definitions in your notes and compare them.
3. In your own words distinguish between the test statistic and the critical value.
4. Can the p-value associated with the test statistic ever be zero? Explain.
5. Since the p-value is associated with the test statistic and the null hypothesis is true, what value must the p-value be for us to provisionally reject the null hypothesis?
6. All of our discussions have been about testing the null hypothesis, about accepting or rejecting, provisionally, the null hypothesis. If we reject the null hypothesis, can we say that we have evidence for the alternate hypothesis?
7. What are the p-values for -logP of 5, 6, 7, 8, 9, and 10? Complete the p-value -logP transform table.
8. Instead of $\log_{10}$ transform, create a similar table but for negative natural log transform. Which is more convenient? Hint: `log(x, base=exp(1))`

8.2: The controversy over proper hypothesis testing

Introduction

Over the next several chapters we will introduce and develop an approach to statistical inference, which has been given the title "Null Hypothesis Significance Testing" or NHST.

In outline, NHST proceeds with

- statements of two hypotheses, a **null hypothesis**, H_O, and an **alternate hypothesis**, H_A
- calculate a **test statistic** comparison of the null hypothesis (assuming some characteristic of data).
- The value of the test statistic is to be compared to a **critical value** for the test, identified for the assumed **probability distribution** at associated **degrees of freedom** for the statistical function, and assigned **Type I error rate**.

We will expand on these statements later in this chapter, so stay with me here. Basically, the null hypothesis is often a statement like "the responses of subjects from the treatment and control groups are the same", e.g., no treatment effect. Note that the alternate hypothesis, e.g., hypertensive patients receiving hydalazine for six weeks have lower systolic blood pressure than patients receiving a placebo (Campbell et al 2011), would be the *scientific hypothesis* we are most interested in. But in the Frequentist NHST approach we test the null hypothesis, not the alternate hypothesis. This framework over proper hypothesis testing is the basis of the Bayesian vs Frequentist controversy.

Consider the independent sample t-test (see Chapter 8.5 and 8.6), our first example of a **parametric test**.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

After plugging in the sample means and the standard error for the difference between the means, we calculate t, the test statistic of the t-test. The critical value is treated as a cut-off value in the NHST approach. We have to set our Type I error rate before we start the experiment, and we have available the degrees of freedom for the test, which follows from the sample size. With these in hand, the critical value is found by looking in the t-table of probabilities (or better, use R).

For example, what is the critical value of a t-test with 10 degrees of freedom and Type I error of 5%?

In Rcmdr, choose **Distributions → Continuous distributions → t distribution → t quantiles…**

Figure 8.2.1: Screenshot of t-quantiles menu in Rcmdr.

Note we want Type I equal to 5%. Since their are two tails for our test, we divide 5% by two and enter 0.025 and select the Upper tail.

R output:

```
> qt(c(.025), df=10, lower.tail=FALSE)
[1] 2.228139
```

which is the same thing we would get if we look up on the t-distribution table (Fig. 8.2.2).

α(1)	0.25	0.1	0.05	0.025	0.01
α(2)	0.5	0.2	0.1	0.05	0.02
DF/1	1.000	3.078	6.314	12.706	31.821
2	0.816	1.886	2.920	4.303	6.965
3	0.765	1.638	2.353	3.182	4.541
4	0.741	1.533	2.132	2.776	3.747
5	0.727	1.476	2.015	2.571	3.365
6	0.718	1.440	1.943	2.447	3.143
7	0.711	1.415	1.895	2.365	2.998
8	0.706	1.397	1.860	2.306	2.896
9	0.703	1.383	1.833	2.262	2.821
10	0.700	1.372	1.812	2.228	2.764

Figure 8.2.2: Screenshot of portion of t-table with highlighted (red) critical value for 10 degrees of freedom.

If the test statistic is greater than the critical value, then the conclusion is that the null hypothesis is to be provisionally rejected. We would like to conclude that the alternative hypothesis should favored as best description of the results. However, we cannot — the **p-value** simply tells us how likely our results would be obtained and if the null hypothesis was true. Confusingly, however, you cannot interpret the p-value as telling you the probability (how likely) that the null hypothesis is true. If, however, the test statistic is less than the critical value, then the conclusion is that the null hypothesis is to be provisionally accepted.

The test statistic can be assigned a probability or p-value. This p-value is judged to be large or small relative to an *a priori* error probability level cut off called the Type I error rate. Thus, NHST as presented in this way may be thought of as a decision path — if the test statistic is greater than the critical value, which will necessarily mean that the p value is less than the Type I error rate, then we make one type of conclusion (reject H_O). In contrast, if the test statistic is less than the critical value, which will mean that the p-value associated with the test statistic will be greater than the Type I error rate, then we conclude something else about the null hypothesis. The various terms used in this description of NHST will be defined in Chapter 8.3.

Sounds confusing, but, you say, OK, what exactly *is* the controversy? The controversy has to do whether the probability or p-value can be interpreted as **evidence** for a hypothesis. In one sense, the smaller the p-value, the stronger the case to reject the null hypothesis, right? However, just because the p-value is small — the event is rare — how much evidence do we have that the null hypothesis is true? Not necessarily, and so we can only conclude that the p-value is one part of what we may need for evidence for or against a hypothesis (hint: part of the solution is to consider **effect size** — introduced in Chapter 9.2 — and the **statistical power of the test**, see Ch 11). What follows was covered by Goodman (1988) and others. Here's the problem. Consider tossing a fair coin ten times, with the resulting trial yielding nine out of ten heads (e.g., a value of one, with tails equal to zero).

R code:

```
set.seed(938291156)
rbinom(10,1,0.5)
[1] 1 1 1 1 1 1 0 1 1 1
```

Note:

To get this result I repeated `rbinom()` a few times until I saw this rare result. I then used the command `get_seed()` from mlr3misc package to retrieve current seed of R's random number generator. Initialize the random seed with the command `set.seed()` .

While rare (binomial probability 0.0098), do we take this as evidence that the coin is not fair? By itself, the p-value provides no information about the alternative hypothesis. More about p-value follows below in sections *What's wrong with the p-value from NHST?* and *The real meaning and interpretation of P-values*.

Statisticians have been aware of limitations of the NHST approach for years (see editorial by Wasserstein et al 2019), but only now is the message getting attention of researchers in the biosciences and other fields. In fact, the New York Times recently had a nice piece by F.D. Flam ("The Odds, Continually Updated," 29 Sep 2014) on the controversy and the **Bayesian** alternative. Like most controversies there are strong voices on either side, and it can be difficult as an outsider to know which position to side with (Fig. 8.2.3).

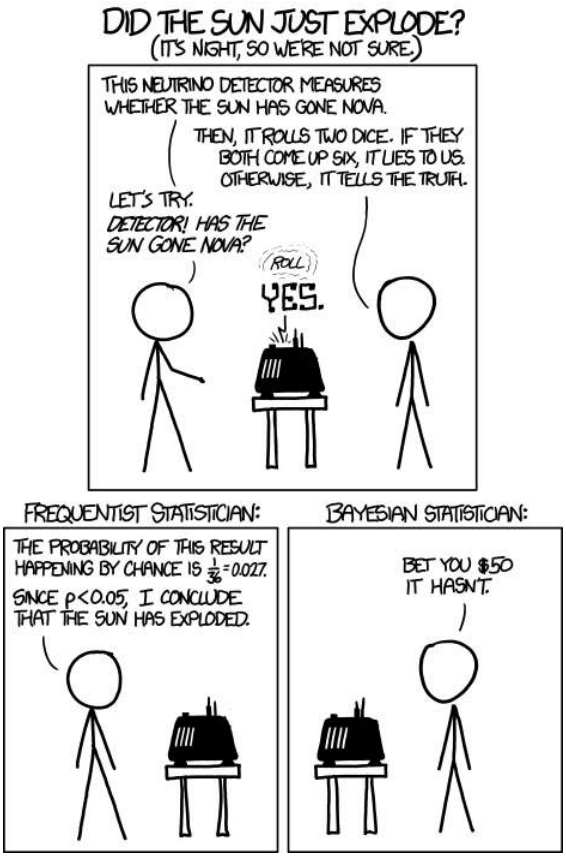


Figure 8.2.3: xkcd: Frequentists vs. Bayesians, https://xkcd.com/1132/.

The short answer is — as you go forward do realize that there is a limitation to the frequentist approach and to be on the correct side of the controversy, you need to understand what you can conclude from statistical results. NHST is by far the most commonly used approached in biosciences (e.g., out of 49 research articles I checked from four randomly selected issues of 2015 *PLoS Biology*, 43 used NHST, 2 used a likelihood approach, none used Bayesian statistics). The NHST is also the overwhelming manner in which we teach introductory statistics courses (e.g., checking out the various MOOC courses at www.coursera.org, all of the courses related to Basic Statistics or Inferential Statistics are taught primarily from the NHST perspective). However, right from the start I want to emphasize the limits of the NHST approach.

If the purpose of science is to increase knowledge, then the NHST approach by itself is an inadequate framework at best, and in the eyes of some, worthless! Now, I think this latter sentiment is way over the top, but there is a need for us to stop before we begin, in effect, to set the ground rules for what can be interpreted from the NHST approach. The critics of NHST have a very important point, and that needs to be emphasized, but we will also defend use and teaching of this approach so that you are not left with the feeling that somehow this is a waste of time or that you are being cheated from learning the latest knowledge on the subject of statistical inference. The controversy hinges on what probability means.

P-values, statistical power, and replicability of research findings

Science, as a way of knowing how the world works, is the only approach that humans have developed that has been empirically demonstrated to work. Note how I narrowed what science is good for — if we are asking questions about the material world, then science should be your toolkit. Some (e.g., Platt 1964), may further argue that there are disciplines in science that have been more successful (e.g., molecular biology) than others (e.g., evolutionary psychology, cf discussion in Ryle 2006) at advancing our knowledge about the material world. However, to the extent that research findings are based solely on statistical results, there is

reason to believe that many studies in fact have not recovered truth (Ioannidis 2005). In a review of genomics, it was reported that findings of gene expression differences by many microarray studies were not reproducible (Allison et al 2006). The consensus is that confidence in the findings should hold only for the most abundant gene transcripts of many microarray gene expression profiling studies, a conclusion that undercuts the perceived power of the technology to discover new causes of disease and the basis for individual differences for complex phenotypes. Note that when we write about failure of research reproducibility we are not including cases of alleged fraud (Carlson 2012 on Duke University oncogenomics case), we are instead highlighting that these kinds of studies often lack statistical power; hence, when repeated, the experiments yield different results.

Frequentist and Bayesian Probabilities

Turns out there is a lot of philosophical problems around the idea of "probability," and three schools of thought. In the **Fisherian approach** to testing, the researcher devises a null hypothesis, H_O, collects the data, then computes a probability (p-value) of the result or outcome of the experiment. If the p-value is small, then this is inferred as little evidence in support of the null hypothesis. In the **Frequentists' approach**, the one we are calling NHST, the researcher devises two hypotheses, the null hypothesis, H_O, and an alternate hypothesis, H_A. The results are collected from the experiment and, prior to testing, a Type I error rate (α, chance) is defined. The Type I error rate is set to some probability and refers to the chance of rejecting the null hypothesis purely due to random chance. The Frequentist then computes a p-value of result of the experiment and applies a decision criterion: If p-value is greater than Type I error rate, then provisionally accept null hypothesis. In both the Fisherian and Frequentist approaches, the probability, again defined at the relative frequency of an event over time, is viewed as a physical, objective and well-defined set of values.

Bayesian approach: based on **Bayes conditional probability**, one identifies the **prior** (subjective) **probability** of an hypothesis, then, adjusts the prior probability (down or up) as new results come in. The adjusted probability is known as **posterior probability** and it is equal to the **likelihood function** for the problem. The posterior probability is related to the prior probability and this function can be summarized by the **Bayes factor** as evidence the evidence against the null hypothesis. And that's what we want, a metric of our evidence for or against the null hypothesis.

> **Note:**
>
> A probability distribution function (PDF) is a function of the sample data and returns how likely that particular point will occur in the sample. The distribution is given. The likelihood function approaches this from a different direction. The likelihood function takes the data set as a given and represents how likely are the different parameters for your distribution.

We can calibrate the Bayesian probability to the frequentist p-value (Selke et al 2001; Goodman 2008; Held 2010; Greenland and Poole 2012). Methods to achieve this calibration vary, but the **Fagan nomogram** proposed by Held (2010) is a good tool for us as we go forward. We can calculate our NHST p-value, but then convert the p-value to a Bayes factor by looking at the nomogram. I mention this here not as part of your to-do list, but rather as a way past the controversy: the NHST p-value can be transcribed to a Bayesian conditional probability.

Likelihood

Before we move on there is one more concept to introduce, that of **likelihood**. We describe a model (an equation) we believe can generate the data we observe. By constructing different models with different parameters (hypotheses), you generate a statistic that yields a **likelihood value**. If the model fits the data, then the likelihood function has a small value. The basic idea then is to compare related, but different models to see which fits the data better. We will use this approach when comparing linear models when we introduce multiple regression models in Chapter 18.

What's wrong with the *p*-value from NHST?

Well, really nothing is "wrong" with the p-value.

Where we tend to get into trouble with the p-value concept is when we try and interpret it. See below, Why is this important to me as a beginning student? The p-value is not evidence for a position, it is a statement about error rates. The p-value from NHST can be viewed as the culmination of a process that is intended to minimize the chance that the statistician makes an error.

In Bayesian terms, the p-value from NHST is the probability that we observe the data (e.g., the differences between two sample means), assuming the null hypothesis is true. If we want to interpret the p-value in terms of evidence for a proposition, then we want the conditional error probability.

Sellke et al (2001) provided a calibration of p-values and, assuming that the prior probabilities of the null hypothesis and the alternative hypothesis are equal (that is, that each have a prior probability of 0.5), by using a formula provided by them (equation 3), we can correct our NHST p-value into a probability that can be interpreted as evidence in favor of the interpretation that the null hypothesis is true. In Bayesian terms, this is called the posterior probability of the null hypothesis. The formula is

$$conditional \ error \ probability = \left\{ 1 + \left(-e \cdot p \cdot \ln \left[p\right]^{-1} \right) \right\}^{-1} \nonumber$$

where e is Euler's number, ln is the base of the natural logarithm, and p is the p-value from the NHST. This calibration works as long as $p < \frac{1}{e}$ (Sellke et al 2001).

By convention we set the Type I error at 5% (cf Cohen 1994). How strong of evidence is a p-value near 5% against the null hypothesis being true, again, under the assumption that the prior probability of the null hypothesis being true is 50%? Using the above formula I constructed a plot of the calculated conditional error probability values against p-values (Fig. 8.2.4).

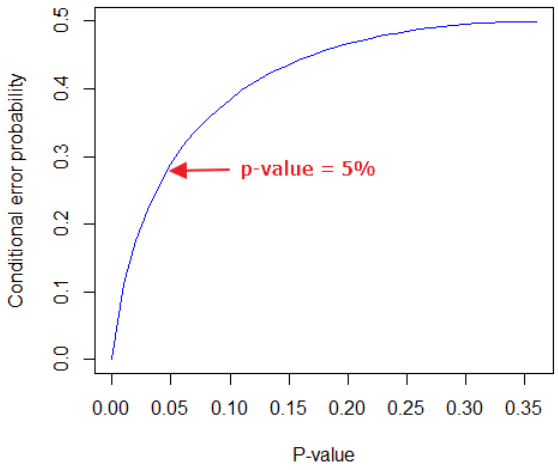


Figure 8.2.4: Conditional error probability values plotted against p-values.

As you can see, a p-value of 5% is not strong evidence at just 0.289. Not until p-values are smaller than 0.004 does the conditional error probability value dip below 0.05, suggesting strong evidence against the null hypothesis being true.

R note: For those of you keeping up with the R work, here's the code for generating this plot. Text after "#" are comments and are not interpreted by R.

At the R prompt type each line:

```
NHSTp = seq(0.00001,0.37,by=0.01)      #create a sequence of numbers between 0.0001 a
CEP = (1+(-1*exp(1)*NHSTp*log(NHSTp))^-1)^-1    #equation 3 from Sellke et al 2001
plot(NHSTp,CEP,xlab="P-value", ylab="Conditional error probability",type="l",col="blu
```

Why is this important to me as a beginning student?

As we go forward I will be making statements about p-values and Type I error rates and null hypotheses and even such things as false positives and false negative. We need to start to grapple with what exactly can be said by p-values in the context of statistical inference, and to recognize that we will sometimes state conclusions that cut some corners when it comes to interpreting p-values. And yet, you (and all consumers of statistics!) are expected to recognize what p-values mean. Always.

The real meaning and interpretation of P-values

This is as good of a time as any to make some clarification about the meaning of p-value and the whole inference concept. Fisher indeed came up with the concept of the p-value, but its use as a decision criterion owes to others and Fisher disagreed strongly with use of the p-value in this way (Fisher 1955; Lehmann 1993).

Here are some common p-value corner-cutting statements to avoid using (after Goodman 2008; Held 2010). P-values are sometimes interpreted, incorrectly, as any of the following:

1. the probability of obtaining the observed data under the assumption of no real effect

2. an observed type-I error rate
3. the false discovery rate, i.e. the probability that a significant finding is a “false positive”
4. the (posterior) probability of the null hypothesis.

So, if p-values don’t mean any of these things, what does a p-value mean? It means that we begin by assuming that there is no effect of our treatments — the p-value is then the chance we will get as large of a result (our test statistic) and the null hypothesis is true. Note that this definition does not include a statement about evidence of the null hypothesis being true. To get evidence of “truth” we need additional tools, like the Bayes Factor and the correction of the p-value to the conditional error probability (see above). Why not dump all of the NHST and go directly to a Bayesian perspective, as some advise? The single best explanation was embedded in the assumption we made about the prior probability in order to calculate the conditional error probability. We assumed the prior probability was 50%. For many, many experiments, that is simply a guess. The truth is we generally don’t know what the prior probability is. Thus, if this assumption is incorrect, then the justification for the formula by Sellke et al (2001) is weakened, and we are no closer to establishing evidence than before. The take-home message is that it is unlikely that a single experiment will provide strong evidence for the truth. Thus the message is repeat your experiments — and you already knew that! And the Bayesians can tell us that the addition of more and more data reduces the effect of the particular value of the prior probability on our calculation of the conditional error probability. So, that’s the key to this controversy over the p-value.

Reporting p-values

Estimated p-values can never be zero. Students may come to use software that may return p-values like “0” — I’m looking at you Google Sheets re: default results from CHISQ.TEST() — but again, this does not mean the probability of the result is zero. The software simply reports values to two significant figures and failed to round. Some journals may recommend that 0 should be replaced by $p < 0.01$ or even < 0.05 inequalities, but the former lacks precision and the latter over-emphasizes the 5% Type I error rate threshold, the “statistical significance” of the result. In general, report p-value to three significant figures and four digits. If a p-value is small, use scientific notation and maintain significant digits. Thus, a p-value of 0.004955794 should be reported as 0.00496 and a p-value of 0.0679 should be reported as 0.0679. Use R’s `signif()` function, for example p-value reported as 6.334e-05, then

```
signif(6.334e-05,3)
[1] 6.33e-05
```

Rounding and significant figures were discussed in Chapter 3.5. See Land and Altman (2015) for guidelines on reporting p-values and other statistical results.

Questions

1. Revisit Figure 8.2.4 again and consider the following hypothesis — the sun will rise tomorrow.
   - If we take the Frequentist position, what would the null hypothesis be?
   - If we take the Bayesian approach, identify the prior probability.
   - Which approach, Bayesian or Frequentist, is a better approach for testing this hypothesis?
2. Consider the pediatrician who, upon receiving a chest X-ray for a child, notes the left lung has a large irregular opaque area in the lower quadrant. Based on the X-ray and other patient symptoms, the doctor diagnoses pneumonia and prescribes a broad-spectrum antibiotic. Is the doctor behaving as a Frequentist or a Bayesian?
3. With the incorrect p-value interpretations listed above in hand, select an article from PLoS Biology, or any of your other favorite research journals, and read how the authors report results of significance testing. Compare the precise wording in the results section against the interpretative phrasing in the discussion section. Do the authors fall into any of the p-value corner-cutting traps?

8.3: Sampling distribution and hypothesis testing

Introduction

Understanding the relationship between **sampling distributions**, **probability distributions**, and hypothesis testing is the crucial concept in the NHST — **Null Hypothesis Significance Testing** — approach to inferential statistics. is crucial, and many introductory text books are excellent here. I will add some here to their discussion, perhaps with a different approach, but the important points to take from the lecture and text are as follows.

Our motivation in conducting research often culminates in the ability (or inability) to make claims like:

1. "Total cholesterol greater than 185 mg/dl increases risk of coronary artery disease."
2. "Average height of US men aged 20 is 70 inches (1.78 m)."
3. "Species of amphibians are disappearing at unprecedented rates."

Lurking beneath these statements of "fact" for populations (just what IS the population for #1, for #2, and for #3?) is the understanding that not ALL members of the population were recorded.

How do we go from our sample to the population we are interested in? Put another way — How good is our sample? We've talked about how "biostatistics" can be generalized as sets of procedures you use to make inferences about what's happening in populations. These procedures include:

- Have an interesting question
- Experimental design (Observational study? Experimental study?)
- Sampling from populations (Random? Haphazard?)
- Hypotheses: H_O and H_A
- Estimate parameters (characterize the population)
- Tests of hypotheses (inferences)

We have control of each of these — we choose what to study, we design experiments to test our hypotheses…We have already introduced these topics (Chapters 6 – 8).

We also obtain estimates of parameters, and inferential statistics applies to how we report our descriptive statistics (Chapter 3). Estimates of parameters like the sample mean and sample standard deviation can be assessed for accuracy and precision (e.g., confidence intervals).

Sampling distribution

Imagine drawing a sample of 30 from a population, calculating the sample mean for a variable (e.g., systolic blood pressure), then calculating a second sample mean after drawing a new sample of 30 from the same population. Repeat, accumulating one estimate of the mean, over and over again. What will be the shape of this distribution of sample means? The **Central Limit Theorem** states that the shape will be a normal distribution, regardless of whether or not the population distribution was normal, as long as the sample size is large (i.e., **Law of Large Numbers**). We alluded to this concept when we introduced discrete and continuous distributions (Chapter 6).

It's this result from theoretical statistics that allows us to calculate the probability of an event from a sample without actually carrying out repeated sampling or measuring the entire population.

A worked example

To demonstrate the CLT, we want R to help us generate many samples from a particular distribution and calculate the same statistic on each sample. We could make a for loop, but the `replicate()` function provides a simpler framework. We'll sample from the chi-square distribution. You should extend this example to other distributions on your own; see Question 5 below.

Note:

This example is much simpler to enter and run code in the script window, adjusting code directly as needed. If you wish to try to run this through Rcmdr, you'll need to take a number of steps, and likely need to adjust the code and rerun anyway. Some of the steps in would be Rcmdr: Distributions → Continuous distributions → Chi-squared distribution → Sample from chi-square distribution…, then running Numerical summaries and saving the output to an object (e.g., `out`), extracting the values from

the object (e.g., `out$Table`, confirm by running command `str(out)` — `str()` is an R utility to display the structure of an object), then testing the object for normality Rcmdr: Statistics → Test of normality, select Shapiro-Wilk, etc.. In other words, sometimes a GUI is a good idea, but in many cases, work with the script!

Generate x replicate samples (e.g., x = 10, 100, 1000, one million) of 30 each from chi-square distribution with one degree of freedom, test the distribution against null hypothesis (assume normal distributed, e.g., Shapiro-Wilk test, see Chapter 13.3), then make a histogram (Chapter 4.2).

```
x.10 <- replicate(10, {
my.mean <- rchisq(30, 1)
mean(my.mean)
})

normalityTest(~x.10, test="shapiro.test")
hist(x.10, col="orange")
```

Result from R:

```
Shapiro-Wilk normality test

data: x.10
W = 0.87016, p-value = 0.1004
```

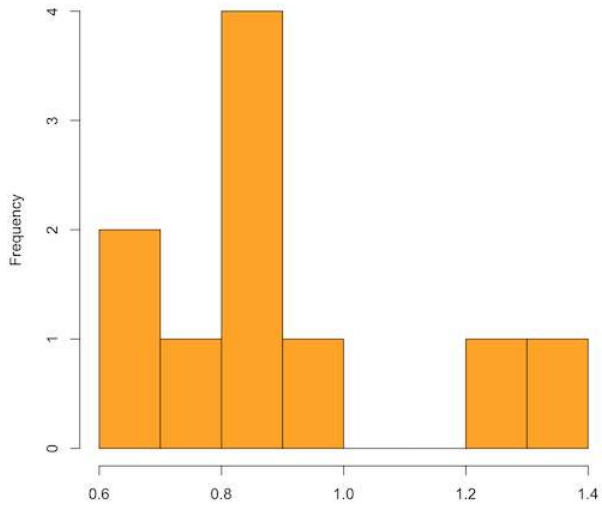


Figure 8.3.1: Means of ten replicate samples drawn at random from chi-square distribution, df = 1.

Modify the code to draw 100 samples, we get:

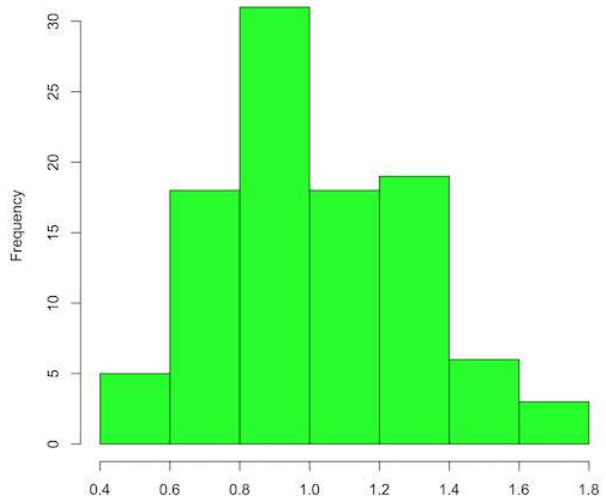


Figure 8.3.2: Means of 100 replicate samples drawn at random from chi-square distribution, df = 1. Results from Shapiro-Wilks test: W = 0.97426, p-value = 0.04721.

And finally, modify the code to draw one million samples, we get:

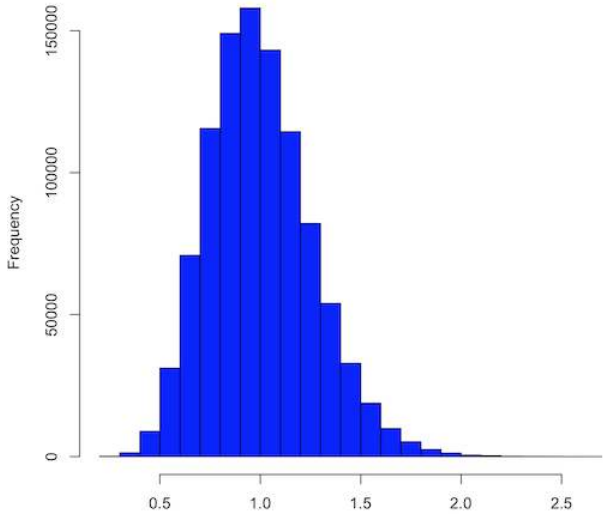


Figure 8.3.1: Means of one million replicate samples drawn at random from chi-square distribution, df = 1. Normality test will fail to run, sample size of 5000 limit.

How to apply sampling distribution to hypothesis testing

First, a reminder of some definitions.

Estimate = we will always (almost) concern ourselves with how good our sample mean (such values are called estimates) is relative to the population mean, the thing we really want, but can only hope to get an estimate of.

Accuracy = how close to the true value is our measure?

Precision = how repeatable is our measure?

How can we tell if we have a good estimate? We want an estimate with an evaluation for accuracy and for precision. The **sampling error** provides an assessment of precision, whereas the **confidence interval** provides a statement of accuracy. We need an estimate of the sampling error for the statistic.

Sample standard error of the mean

We introduced sample error of the mean in section 3.4 of Chapter 3. Everything we measure can have a corresponding statement about how accurate (sampling error) is our estimate! First, we begin by asking, “how accurate is the mean that we estimate from a

sample of a population?" How do we answer this? We could prove it in the mathematical sense of proof (and people have and do) OR we can use the computer to help. We'll try this approach in a minute.

What we will show relates to the standard error of the population mean (SEM) or $s_{\bar{X}}$, whose equation is shown below.

$$SEM = \frac{s^2}{n}$$

Or equivalently, from the standard deviation we have

$$SEM = s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Note that the SEM takes the variance and divides through by the sample size. In general, then, the larger the sample size, the smaller the "error" around the mean. As we work through the different statistical tests, t-tests, analysis of variance, and related, you will notice that the test statistic is calculated as a ratio between a difference or comparison divided by some form of an error measurement. This is to remind you that "everything is variable."

A note on standard deviation (SD) and standard error of the mean (SEM): SD estimates the variability of a sample of X_i values, whereas SEM estimates the variability of a sample of means.

Let's return to our thought problem and see how to demonstrate a solution. First, what is the population? Second, can we get the true population mean?

One way, a direct (but impossible?) approach, would be to measure it — get all of the individuals in a population and measure them, then calculate the population mean. Then, we could compare our original sample mean against the true mean and see how close it was. This can be accomplished in some limited cases. For example, the USA conducts a census of her population every ten years, a procedure which costs billions of dollars. We can then compare samples from the different states or counties to the USA mean. And these statistics are indeed available via the census.gov website. But even the census uses sampling — individuals are randomly selected to answer more questions and from this sample trends in the population are inferred.

So, sampling from populations is the way to go for most questions we will encounter. The procedures we will use to show how a sample mean relates to the population mean are general and may be used to show how any estimate of a **variable** (sample mean and sample standard deviation, etc.), relates to properties of a **parameter**. We'll get to the other issues, but for now, think about sample size.

Sampling from populations is necessary and inevitable, and, to a certain extent, under your control. But how many individuals do we need? The quick answer is for me to direct your attention to the equation for the SEM. Can you see in that ratio the secret to obtaining more precise estimates? There are many ways to approach this question, but let's use the tools from last time, those based on properties of a normal distribution.

If we can view the sampling as having come from a population at least approximately normally distributed for our variable, then we can now examine empirically the effect of different sample sizes on the estimate of the mean.

A hint: variability is important!

From one population we obtain two samples, A and B. Sample sizes are

Group A, $n = 9$
Group B, $n = 50$

Assume for now that we know the true mean (μ) and standard deviation (σ) for the population. Note. This is one of the points of why we use computer simulation so much to teach statistics — it allows us to specify what the truth is, then see how our statistical tools work or how our assumptions affect our statistically based conclusions.

$\mu = 47.0$ mm
$\sigma = 12.0$ mm

Confidence intervals

Reliability is another word for **precision**. We define a confidence interval as a statistic to report the reliability of our estimated statistic. We introduced confidence interval in Section 3.4. At least in principle, confidence intervals can be calculated for all

statistics (mean, variance, etc.,) and for all data types. Confidence intervals define a **lower limit**, L, and an **upper limit**, U, and that you are making a statement that you are "95% certain that the true value (parameter value) is between these two limits."

We previously reported how to calculate an approximate confidence intervals for proportions and for NNT; simply multiple standard error estimate by 2. Here we introduce an improved approximate calculation of the 95% confidence interval for the sample mean:

$$CI\ 95\% = \bar{X} \pm Z \cdot s_{\bar{X}}$$

where Z is something you would look up from the table of the normal distribution. For a 95% confidence interval, $100\% - 95\% = 5\%$ and divide 5% by two: the lower limit corresponds to 2.5% and the upper limit corresponds to 2.5% on our normal distribution. We look up the table and we find that Z for 0.025 is 1.96, and that is the value we would plug into our equation above. For large sample sizes, you can get a pretty decent estimate of the confidence interval by replacing 1.96 with "2."

Questions

1. What is the probability of having a sample mean *greater* than 50 (mean > 50) for a sample of n = 9 ?

We'll use a slight modification of the Z-score equation we introduced in Chapter 6.6 — the modification here is that previously we referred to the distribution of X_i values and how likely a particular observation would be. Instead, we can use the Z score with the standard normal distribution (aka Z-distribution), approach to solving how likely an estimated sample mean is given the population parameters μ and σ. Recall the Z score:

$$Z = \frac{X_i - \mu}{\sigma}$$

We have everything we need except the SEM, which we can calculate by dividing the standard deviation by squared root of sample size.

For $\bar{X} = 50$, $\sigma = 12.0$ (given above), $\mu = 47$, and $n = 9$, plug in the values:

$$s_{\bar{X}} = \frac{12.0}{\sqrt{3}} = 4$$

Therefore, after applying the equation for Z score, $Z = 0.75$. This corresponds to how far away from the standard mean of zero.

Look up $Z = 0.75$ from the table of normal distribution. The answer is 0.22663, which corresponds to Z being EQUAL to or GREATER than 0.75, which is what we wanted. Translated, this implies that, given the level of variability in the sample, 22.66% of your sample means would be greater than 50! We write: $P(X > 50.0) = P(Z > 0.75) = 0.2266$.

Some care needs to be taken when reading these tables — make sure you understand how the direction (less than, greater than) away from the mean is tabulated.

2. Instead of *greater*, how would you get the probability *less* than 50?

Total area under the curve is 1 (100%), so subtract 1 from 0.22663, which equals 0.7734.

I recommend that you do these by hand first, then check your answers. You'll need to be able to do this for exams.

Here's how to use `Rcmdr` to do these kind of problems.

Rcmdr: Distributions → Continuous distributions → Normal distribution → Normal probabilities …

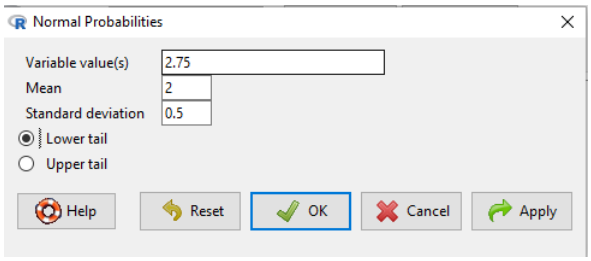


Figure 8.3.4: Screenshot Rcmdr menu to get normal probability.

Here's the answer from Rcmdr:

```
pnorm(c(50), mean=47, sd=12, lower.tail=TRUE)

[1] 0.5987063
```

3. Now, try a larger sample size. For $n = 50$, what is the probability of having a sample mean greater than 50 (mean > 50)?

$\bar{X} = 50$, $\mu = 47$, $\sigma = 12$, $n = 50$, and $SEM = \frac{12.0}{\sqrt{50}} = 1.697$.

Therefore, after applying the equation for Z score, $Z = 1.768$. Look up $Z = 1.768$ (Normal table, subtract answer from 1) and we get 0.0384. This means that 3.84% of your sample means would be greater than 50! We write: $P(X > 50.0) = P(Z > 1.768) = 0.0384$.

Said another way: If you have a sample size of 50 $(N = 50)$ and you obtain a mean greater than 50, then there is only a 3.84% chance that the TRUE MEAN IS 47.

4. What happens if the variability is smaller? Chance σ from 12 to 6, then repeat questions 1 and 4.

5. Repeat the demonstration of Central Limit Theorem and Law of Large Numbers for discrete distributions:

A. **binomial distribution**. Replace `rchisq()` with `rbinom(n, size, prob)` in the `replicate()` function example. See Chapter 6.5

B. **poisson distribution**. Replace `rchisq()` with `rpois(n, lambda)` in the `replicate()` function example. See Chapter 6.5

8.4: Tails of a test

Introduction

The basics of **statistical inference** is to establish the **null** and **alternative hypotheses**. Starting with the simplest cases, where there is one sample of observations and the comparison is against a population (theory) mean, how many possible comparisons can be made? The next simplest is the two-sample case, where we have two sets of observations and the comparison is against the two groups. Again, how many total comparisons may be made?

Let $\bar{X}$, "X bar", equal the sample mean and μ, "mu", represent the population mean. For sample means, designate groups by a subscript, 1 or 2. We then have Table 8.4.1.

Table 8.4.1. Possible hypothesis involving one or two groups.

Comparison	One-sample	Two-sample
1.	$\bar{X} = \mu$	$\bar{X}_1 = \bar{X}_2$
2.	$\bar{X} \neq \mu$	$\bar{X}_1 \neq \bar{X}_2$
3.	$\bar{X} \geq \mu$	$\bar{X}_1 \geq \bar{X}_2$
4.	$\bar{X} \leq \mu$	$\bar{X}_1 \leq \bar{X}_2$
5.	$\bar{X} > \mu$	$\bar{X}_1 > \bar{X}_2$
6.	$\bar{X} < \mu$	$\bar{X}_1 < \bar{X}_2$

Classical statistics classifies inference into null hypothesis, H_O, vs. alternate hypotheses, H_A, and specifies that we test null hypotheses based on the value of the estimated test statistic (see discussion about critical value and **p-value**, Chapter 8.2). From the list of six possible comparisons we can divide them into **one-tailed** and **two-tailed** differences (Table 8.4.1). By "tail" we are referring to the ends or tails of a distribution (Figure 8.4.1, Figure 8.4.2); where do our results fall on the distribution?

Two-tailed hypotheses: Comparison 1 and comparison 2 in the table above are two-tailed hypotheses. We don't ask about the direction of any difference (less than or greater than).

Figure 8.4.1 shows the "two-tailed" distribution — if our results fall to the left (≤ -1.96) or to the right $(\geq +1.96)$ we reject the null hypothesis (blue regions in the curve). We divide the type I error into two equal halves.

Note:

It's a nice trick to shade in **regions of the curve**. A package `tigerstats` includes the function `pnormGC` that simplifies this task.

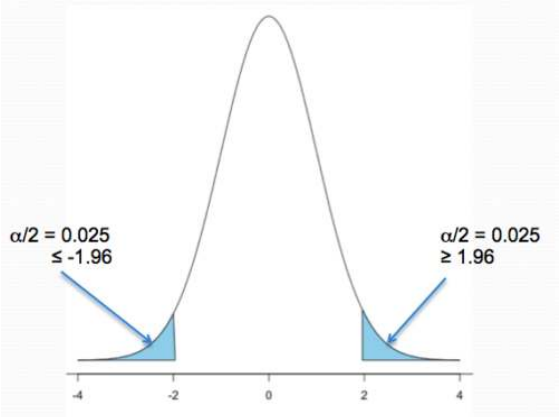


Figure 8.4.1: Two-tailed distribution.

```
library(tigerstats)
pnormGC(c(-1.96,1.96), region="outside", mean=0, sd=1,graph=TRUE)
```

Figure 8.4.2 shows the "one-tailed" distribution — if our alternate hypothesis was that the sample mean was less than the population mean, then our fall to the left (≤ -1.645) for the "lower tail" of the distribution. If, however, our alternate hypothesis was that the sample mean was

greater than the population mean, then our region of interest falls to the right ($\geq +1.645$). Again, we reject the null hypothesis (blue regions in the curve). Note for one-tailed hypothesis, all **Type I error** occurs in the one area, not both, so α (alpha) remains 0.05 over the entire rejection region.

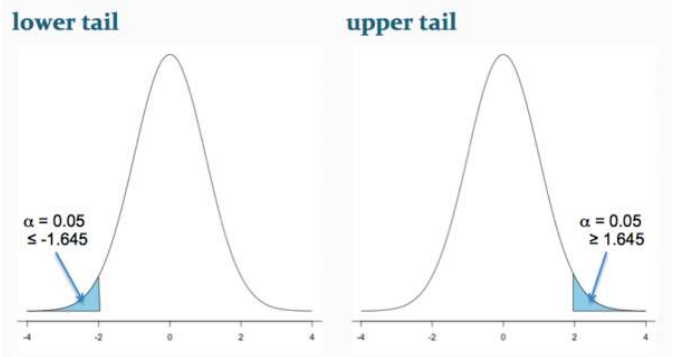


Figure 8.4.2: One-tailed distribution, lower tail (left) and upper tail (right).

```
library(tigerstats)
pnormGC(1.645, region="above", mean=0, sd=1,graph=TRUE)
pnormGC(-1.645, region="below", mean=0, sd=1,graph=TRUE)
```

One-tailed hypotheses: Comparison 3 through comparison 6 in the table are one-tailed hypotheses. The direction of the difference matters.

Note:

A simple trick to writing one-tailed hypotheses: first write the alternate hypothesis because the null hypothesis includes all of the other possible outcomes of the test.

Examples

Let's consider some examples. We learn best by working through cases.

Chemotherapy as an approach to treat cancers owes its origins to the work of Dr. Sidney Farber, among others in the 1930s and '40s (DeVita and Chu 2008; Mukherjee 2011). Following up on the observations of others that folic acid (vitamin B_9) improved anemia, Dr Farber believed that folic acid might reverse the course of leukemia (Mukherjee 2011). In 1946 he recruited several children with acute lymphoblastic leukemia and injected them with folic acid. Instead of ameliorating their symptoms (e.g., white blood cell counts and percentage of abnormal immature white blood cells, called blast cells), treatments accelerated progression of the disease. That's a scientific euphemism for the reality — the children died sooner in Dr. Faber's trial than patients not enrolled in his study. He stopped the trials. Clearly, adding folic acid was not a treatment against this leukemia.

Question 1. Do you think these experiments are one-sample or two-sample? Hint: Is there mention of a control group?

Answer: There's no mention of a control group, but instead, Dr. Faber would have had plenty of information about the progression of this disease in children. This was a one-sample test.

Question 2. What would be a reasonable interpretation of Dr. Faber's alternate hypothesis with respect to percentage of blast cells in patients given folic acid treatment? Your options are

1. Folic acid supplementation has an effect on blast counts.
2. Folic acid supplementation reduces blast counts.
3. Folic acid supplementation increases blast counts.
4. Folic acid supplementation has no effect on blast counts.

Answer: At the start of the trials, it is pretty clear that the alternate hypothesis was intended to be a one-tailed test (option 2). Dr. Faber's alternative hypothesis clearly was that he believed that addition of folic acid would reduce blast cell counts. However, that they stopped the trials shows that they recognized that the converse had occurred, that blast counts increased; this means that, from a statistician's point of view, Dr Faber's team was testing a two-sided hypothesis (option 1).

Here's another example.

Dr. Farber reasoned that if folic acid accelerated leukemia progression, perhaps anti-folic compounds might inhibit leukemia progression. Dr Farber's team recruited patients with acute lymphoblastic leukemia and injected them with a folic acid agonist called aminopterin. Again, he predicted that blast counts would reduce following administration of the chemical. This time, and for the first time in recorded medicine, blast

counts of many patients drastically reduced to normal levels and the patients experienced remissions. The remissions were not long-lasting and all patients eventually succumbed to leukemia. Nevertheless, these were landmark findings — for the first time a chemical treatment was shown to significantly reduce blast cell counts, even leading to remission, if however brief (Mukherjee 2011).

Try Question 3 and Question 4 yourself.

Question 3. Do you think these experiments are one-sample or two-sample? Hint: Is there mention of a control group?

Question 4. What would be a reasonable interpretation of Dr Faber's alternate hypothesis with respect to percentage of blast cells in patients given aminopterin treatment? Your options are

1. Aminopterin supplementation has an effect on blast counts.
2. Aminopterin supplementation reduces blast counts.
3. Aminopterin supplementation increases blast counts.
4. Aminopterin supplementation has no effect on blast counts.

Pros and Cons to One-sided testing

Here's something to consider: why not restrict yourself to one-tailed hypotheses? Here's the pro-argument. Strictly speaking you gain statistical power to test the null hypothesis. For example, look up the t-test distribution for degrees of freedom equal to 20 and compare $\alpha_{(1)}$ (one tail) vs. $\alpha_{(2)}$ (two-tail). You will find that for the one-tailed test, the critical value of the t-distribution with $df = 20$ is 1.725, whereas for the two-tailed test, the critical value of the t-distribution with the same df numbers is 2.086. Thus, the difference between means can be much smaller in the one-tailed test and prove to be "statistically significant." Put simply, with the same data, we will reject the Null Hypothesis more often with one-tailed tests.

The con-argument. If you use a one-tailed test you MUST CLEARLY justify its use and be aware that a deviation in the opposite direction MUST be ignored! More specifically, you interpret a one-tailed result in the opposite direction as acceptance of the null — you cannot, after the fact, change your mind and start speaking about "statistically significant differences" if you had specified a one-tailed hypothesis and the results showed differences in the opposite direction.

Note:

Recall also that, by itself, statistical significance judged by the p-value against a specified cut-off critical value is not enough to say there is evidence for or against the hypothesis. For that we need to consider effect size, see Power analysis in Chapter 11.

Questions

1. For a Type I error rate of 5% and the following degrees of freedom, compare the critical values for one tail test and a two tailed test of the null hypothesis.

- $df = 5$
- $df = 10$
- $df = 15$
- $df = 20$
- $df = 25$
- $df = 30$

2. Using your findings from Question 1, make a scatterplot with degrees of freedom on the horizontal axis and critical values on the vertical axis. What trend do you see for the difference between one- and two-tailed tests as degrees of freedom increase?
3. A clinical nutrition researcher wishes to test the hypothesis that a vegan diet lowers total serum cholesterol levels compared to an omnivorous diet. What kind of hypothesis should he use, one-tailed or two-tailed? Justify your choice.
4. Spironolactone, introduced in 1953, is used to block aldosterone in hypertensive patients. A newer drug eplerenone, approved by the FDA in 2002, is reported to have the same benefits as spironolactone (reduced mortality, fewer hospitalization events), but with fewer side effects compared with spironolactone. Does this sentence suggest a one-tailed test or a two-tailed test?
5. Write out the appropriate null and alternative hypothesis statements for the spironolactone and eplerenone scenario.
6. You open up a bag of Original Skittles and count the number of green, orange, purple, red, and yellow candies in the bag. What kind of hypothesis should be used, one-tailed or two-tailed? Justify your choice.
7. Verify the probability values from the table of standard normal distribution for Z equal to -1.96, -1.645, 1.645, and 1.96.

8.5: One sample t-test

Introduction

We're now talking about the traditional, classical two-group comparison involving continuous data types. Thus begins your introduction to **parametric statistics**. One-sample tests involve questions like "how many (what proportion of) people would we expect are shorter or taller than two standard deviations from the mean?" This type of question assumes a population and we use properties of the normal distribution and, hence, these are called parametric tests because the assumption is that the data has been sampled from a particular probability distribution.

However, when we start asking questions about a **sample statistic** (e.g., the **sample mean**), we cannot use the **normal distribution** directly, i.e., we cannot use Z and the **normal table** as we did before (Chapter 6.7). This is because we do not know the population standard deviation and therefore must use an estimate of the variation (s) to calculate the standard error of the mean.

With the introduction of the t-statistic, we're now into full inferential statistics-mode. What we do have are estimates of these parameters. The t-test — aka **Student's t-test** — was developed for the purpose of testing sample means when the true population parameters are not known.

> **Note:**
>
> It's called Student's t-test after the pseudonym used by William Gosset.

This is the equation of the **one sample t-test**. Note the resemblance in form with the **Z-score**!

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

where $s_{\bar{X}}$ is the **sample standard error of the sample mean** (SEM).

For example, weight change of mice given a hormone (leptin) or **placebo**. $\bar{X} = 5$ g, but under the null hypothesis, the mean change is "really" zero $\mu = 0$. How unlikely is our value of 5 grams?

> **Note:**
>
> Notice how I snuck in "placebo" and mice? Do you think the concept of placebo is appropriate for research with mice, or should we simply refer to it as a **control treatment**? See Ch. 5.4 – Clinical trials for review.

Speaking of **null hypotheses**, can you say (or write) the null and alternative hypotheses in this example? How about in symbolic form?

We want to know if our sample mean could have been obtained by chance alone from a population where the true change in weight was zero.

$s = 3$, $n = 20$, and $s_X = \frac{s}{\sqrt{n}} = \frac{3}{\sqrt{20}} = 0.6708$

We take these values and plug them into our equation of the t-test:

$$t = \frac{5 - 0}{0.67} = 7.45$$

Then recall that **Degrees of Freedom** are $DF = n - 1$, so we have $DF = 20 - 1 = 19$ for the one sample t-test. And the **Critical Value** is found in the appropriate table of critical values for the t distribution (Fig. 8.5.1).

α(1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α(2)	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
DF/1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	
5	0.727	1.476	2.015	2.571	3.365	4.032			
6	0.718	1.440	1.943					3.733	4.073
7					2.583	2.921	3.252	3.686	4.015
	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792

Figure 8.5.1: Table of a portion of the Critical values of the t distribution. Red selections highlight critical value for t-test at $\alpha = 5\%$ and $df = 19$.

Note:

See our table of critical values of t distribution.

Or, and better, use R

```
qt(c(0.025), df=19, lower.tail=FALSE)
```

where `qt()` is function call to find t-score of the p[th] percentile (cf 3.3 – Measures of dispersion) of the Student t distribution. For a two-tailed test, we recall that 0.025 is lower tail and 0.025 is upper tail.

In this example we would be willing to reject the Null Hypothesis if there was a positive OR a negative change in weight.

This was an example of a "two-tailed test," which is "2-tail" or $\alpha_{(2)}$ in the table of critical values of the t distribution.

The critical value for $\alpha_{(2)} = 0.05$ and $df = 19$ is 2.093. Do we accept or reject the Null Hypothesis?

A typical inference workflow

Note the general form of how the statistical test is processed, a form which actually applies to any statistical inference test.

1. Identify the type of data
2. State the null hypothesis (2-tailed? 1-tailed?)
3. Select the test statistic (t-test) and determine its properties
4. Calculate the test statistic (the value of the result of the t-test)
5. Find degrees of freedom
6. For the DF, get the critical value
7. Compare critical value to test statistic
8. Do we accept or reject the null hypothesis?

And then we ask, given the results of the test of inference, **What is the *biological* interpretation?** Statistical significance is not necessarily evidence of biological importance. In addition to statistical significance, the magnitude of the difference — the **effect size** — is important as part of interpreting results from an experiment. Statistical significance is at least in part because of sample size — the large the sample size, the smaller the standard error of the mean, therefore even small differences may be statistically significant, yet biologically unimportant. Effect size is discussed in Ch. 9.1 – Chi-square test: Goodness of fit, Ch. 11.4 – Two-sample effect size and Ch. 12.5 – Effect size for ANOVA.

R Code

Let's try a one-sample t-test. Consider the following data set: body mass of four geckos and four Anoles lizards (*Dohm unpublished data*).

For starters, let's say that you have reason to believe that the true mean for all small lizards is 5 grams (g).

```
Geckos: 3.186, 2.427, 4.031, 1.995
Anoles: 5.515, 5.659, 6.739, 3.184
```

Get the data into R (Rcmdr)

By now you should be able to load this data in one of several ways. If you haven't already entered the data, check out Part 07. Working with your own data in Mike's Workbook for Biostatistics.

Once we have our **data.frame**, proceed to carry out the statistical test.

To get the one-sample t-test in `Rcmdr`, click on **Statistics → Means → Single-sample t-test…** Because there is only one numerical variable, `Body.mass`, that is the only one that shows up in the Variable (pick one) window (Fig. 8.5.2)

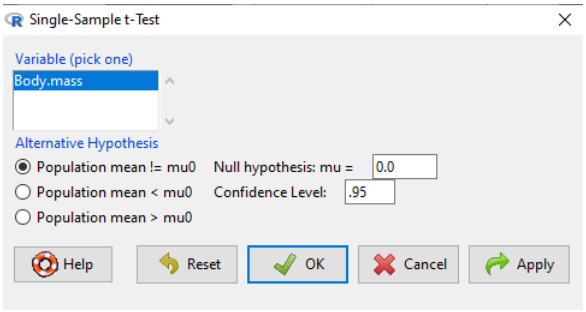


Figure 8.5.2: Screenshot of Rcmdr single-sample t-test menu.

Type in the value 5.0 in the Null hypothesis: mu = box.

Question 3: Quick! Can you write, in plain old English, the statistical null hypothesis???

Click OK

The results go to the Output Window.

```
t.test(lizards$Body.mass, alternative = 'two.sided', mu = 5.0, conf.level = .95)
```

```
One Sample t-test
data: lizards$Body.mass
t = -1.5079, df = 7, p-value = 0.1753
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
2.668108 5.515892
sample estimates:
mean of x 4.092
```

Let's identify the parts of the R output from the one sample t-test. R reports the name of the test and identifies:

1. The dataset variable used (`lizards$Body.mass`). The data set was called "lizards" and the variable was "Body.mass". R uses the dollar sign ($) to denote the data set and variable within the data set.
2. The value of the statistic was $t = -1.5079$. It is negative because the sample mean was less than the population mean — you should be able to verify this!
3. The degrees of freedom: $df = 7$
4. The p-value = 0.1753
5. Confidence level = 95%
6. The sample mean = 4.092

Take a step back and review

Let's make sure we "get" the logic of the hypothesis testing we have just completed. Consider the one-sample t-test.

Step 1. Define H_O and H_A. The null hypothesis might be that a sample mean equals 5. $H_O : \bar{X} = 5$

The alternate is that the sample mean is not equal to 5. $H_A : \bar{X} \neq 5$

Where did the value 5 come from? It could be a value from the literature (does the new sample differ from values obtained in another lab?). The point is that the value is known in advance, before the experiment is conducted, and that makes it a one-sample t-test.

One-tailed hypothesis or two?

We introduced you to the idea of "tails of a test" (Ch. 8.4). As you should recall, a null/alternate hypothesis for a two-tailed test may be written as

$H_O : \bar{X} = \mu$
$H_A : \bar{X} \neq \mu$

Alternatively, we can write a one-tailed test null/alternate hypothesis as

$H_O : \bar{X} < \mu$
$H_A : \bar{X} \geq \mu$

Question 4: Are all possible outcomes of the one-tailed test covered by these hypotheses?

Question 5: What is the SEM for this problem?

Question 6: What is the difference between a one-sample t-test and a one-sided t-test?

Question 7: What are some other possible hypotheses that can be tested from this simple example of two lizard species?

Step 2. Decide how certain you wish to be (with what probability) that the sample mean is different from 5. As stated previously, in biology we say that we are willing to be incorrect.

Step 3. Carry out the calculation of the test statistic. In other words, get the value of t from the equation above by hand, or, if using R (yes!) simply identify the test statistic value from the R output after conducting the one-sample t-test.

Step 4. Evaluate the result of the test. If the value of the test statistic is greater than the critical value for the test, then you conclude that the chance (the P-value) that the result could be from that population is not likely and you therefore reject the null hypothesis.

Question 8: What is the critical value for a one-sample t-test with $df = 7$? Hint: you need the table, or better, R: **Rcmdr: Distributions → Continuous distributions → t distributions → t quantiles**. You also need to know three additional things to answer this question.

1. You need to know α, which we have said is generally set at 5.
2. You also need to know the degrees of freedom (DF) for the test. For a one-sample test, $DF = n - 1$, where n is the sample size.
3. You also must know whether your test is one- or two-tailed.

You then use the t-distribution (the tables of the t-distribution at the end of your book) to obtain the critical value. Note that if you use R the actual p-value is returned.

Why learn the equations when I can just do this in R?

`Rcmdr` does this for you for you as soon as you click OK. Rcmdr returns the value of the test statistic and the p-value. R does not show you the critical value, but instead returns the probability that your test statistic is as large as it is AND the null hypothesis is true.

The simple answer is that in order to understand the R output properly you need to know where each item of the output for a particular test comes from and how to interpret it. Thus, the best way is to have the equations available and to understand the algorithmic approach to stastical inference.

Also, this is as good of a time as any to show you how to skip the Rcmdr GUI and go straight to R.

First, create your variables. At the R prompt enter the first variable:

```
liz <- c("G", "G", "G", "G", "A", "A", "A", "A")
```

and then create the second variable:

```
bm <- c(3.186, 2.427, 4.031, 1.995, 5.515, 5.659, 6.739, 3.184)
```

Next, create a data frame. Think of a data frame as another word for worksheet.

```
lizz <- data.frame(liz, bm)
```

Verify that entries are correct. At the R prompt type “lizz” without the quotes and you should see

```
   liz       bm
1    G    3.186
2    G    2.427
3    G    4.031
4    G    1.995
5    A    5.515
6    A    5.659
7    A    6.739
8    A    3.184
```

Carry out the t-test by typing the following at the R prompt:

```
t.test(lizz, bm, alternative='two-sided', mu=5, conf.level=.95)
```

And, like the Rcmdr output, we have for the one-sample t-test the following R output:

```
One Sample t-test
data: lizards$Body.mass
t = -1.5079, df = 7, p-value = 0.1753 alternative hypothesis: true mean is not equal
95 percent confidence interval:
2.668108 5.515892
sample estimates:
mean of x
4.092
```

End of R output

which, as you probably guessed, is the same as what we got from RCmdr.

Question 9: From the R output of the one sample t-test, what was the value of the test statistic?

A. -1.5079
B. 7
C. 0.1753
D. 2.668108
E. 5.515892
F. 4.092

Note. On an exam you will be given portions of statistical tables and output from R. Thus you should be able to evaluate statistical inference questions by completing the missing information. For example, if I give you a test statistic value, whether the test is one-

or two-tailed, degrees of freedom, and the Type I error rate alpha, you should know that you would need to find the critical value from the appropriate statistical table. On the other hand, if I give you R output, you should know that the p-value and whether it is less than the Type I error rate of alpha would be all that you need to answer the question.

Think of this as a basic skill.

In statistics and for some statistical tests, Rcmdr and other software may not provide the information needed to decide that your test statistic is large, and a table in a statistics book is the best way to evaluate the test.

For now, double check Rcmdr by looking up the critical value from the t-table.

Check critical value against our test statistic

$Df = 8 - 1 = 7$

The test is two-tailed, therefore $\alpha(_2)$.

$\alpha = 0.05$ (note that two-tailed critical value is 2.365. t was equal to 1.51 (since t-distribution is symmetrical, we can ignore the negative sign), which is smaller than 2.365 and so we would agree with Rcmdr — we cannot reject the null hypothesis.

Question 10: From the R output of the one sample t-test, what was the P-value?

A. -1.5079
B. 7
C. 0.1753
D. 2.668108
E. 5.515892

Question 11: We would reject the null hypothesis

A. False
B. True

Questions

Eleven questions were provided for you within the text in this chapter. Here's one more.

Question 12. Here's a small data set for you to try your hand at the one-sample t-test and Rcmdr. The dataset contains cell counts, five counts of the numbers of beads in a liquid with an automated cell counter (Scepter, Millipore USA). The true value is 200,000 beads per milliliter fluid; the manufacturer claims that the Scepter is accurate within 15%. Does the data conform to the expectations of the manufacturer? Write a hypothesis then test your hypothesis with the one-sample t-test. Here's the data.

Scepter
258900
230300
107700
152000
136400

8.6: Confidence limits for the estimate of population mean

Introduction

In Chapter 3.4 and Chapter 8.3, we introduced the concept of providing a confidence interval for estimates. We gave a calculation for an approximate confidence interval for proportions and for the Number Needed to Treat (Chapter 7.3). Even an approximate confidence interval gives the reader a range of possible values of a population parameter from a sample of observations.

In this chapter we review and expand how to calculate the **confidence interval for a sample mean**, $\bar{X}$. Because $\bar{X}$ is derived from a sample of observations, we use the ***t*-distribution** to calculate the confidence interval. Note that if the population was known (population standard deviation), then you would use normal distribution. This was the basis for our recommendation to adjust your very approximate estimate of a confidence interval for an estimate by replacing the "2" with "1.96" when you multiply the **standard error of the estimate** (SE) in the equation estimate $\bar{X} \pm 2 \cdot SE$. As you can imagine, the approximation works for large sample size, but is less useful as sample size decreases.

Consider $\bar{X}$; it is a point estimate of μ, the population mean (a **parameter**). But our estimate of $\bar{X}$ is but one of an infinite number of possible estimates. The confidence interval, however, gives us a way to communicate how reliable our estimate is for the population parameter. A 95% confidence interval, for example, tells the reader that we are willing to say (95% confident) the true value of the parameter is between these two numbers (a lower limit and an upper limit). The point estimate (the sample mean) will of course be included between the two limits.

Instead of 95% confidence, we could calculate intervals for 99%. Since 99% is greater than 95%, we would communicate our certainty of our estimate.

> **Note**
>
> Again, the caveats about p-value extend to confidence intervals. See Chapter 8.2.

Question 1: For 99% confidence interval, the lower limit would be smaller than the lower limit for a 95% confidence interval.

A. True ← Answer
B. False

When we set the Type I error rate, α (alpha) = 0.05 (5%), that means that 5% of all possible sample means from a population with mean, μ, will result in t values that are larger than $+t_{0.05(2),df}$ OR smaller than $-t_{0.05(2),df}$.

Why the *t*-distribution?

We use the t-test because, technically, we have a limited sample size and the t-distribution is more accurate than the normal distribution for small samples. Note that as sample size increases, the t-distribution is not distinguishable from the **normal distribution** and we could use ± 1.96 (Fig. 8.6.1).

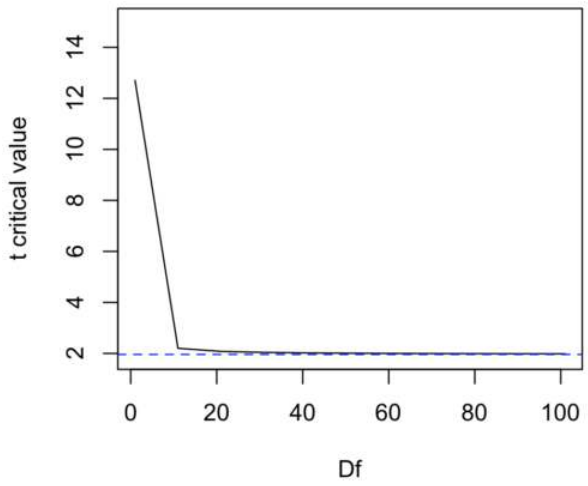


Figure 8.6.1: Copy and Paste Caption here. (Copyright; author via source)

Here's the equation for calculating the confidence interval based on the t-distribution. These set the limits around our estimate of the sample mean. Together, they're called the 95% confidence interval, $1 - 0.05 = 0.95$.

$$P\left[-t_{0.05(2),df} \leq \frac{\bar{X} - \mu}{s_{\bar{X}}} \leq +t_{0.05(2),df}\right] = 0.95$$

Here's a simplified version of the same thing, but generalized to any Type I level…

$$\bar{X} \pm t_{\alpha/2,v} \cdot s_{\bar{X}}$$

This statistic allows us to say that we are 95% confident that the interval $\pm t_{0.05(2),df}$ includes the true value for μ. For this confidence interval you need to identify the critical t value at 5%. Thus, you need to know the degrees of freedom for this problem, which is simply $n - 1$, the sample size minus one.

It is straightforward to calculate these by hand, but…

Set the **Type I error rate**, calculate the **degrees of freedom** (df):

- $n - 1$ samples for one sample test
- $n - 1$ pairs of samples for paired test
- $n - 2$ samples for two independent sample test

and lookup the **critical value** from the t table (or from the t distribution in R). Of course, it is easier to use R.

In R, for the one tail critical value with seven degrees of freedom, type at the R prompt:

```
qt(c(0.05), df=7, lower.tail=FALSE)
[1] 1.894579
```

For the two-tail critical value:

```
qt(c(0.025), df=7, lower.tail=FALSE)
[1] 2.364624
```

Or, if you prefer to use R Commander, then follow the menu prompts to bring up the **t quantiles** function (Fig. 8.6.2 and Fig. 8.6.3).

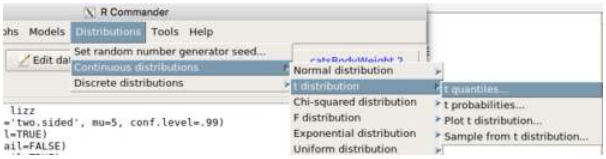


Figure 8.6.2: Drop down menu to get t-distribution.

Note:

Quantiles divide probability distribution into equal parts or intervals. **Quartiles** have four groups, **deciles** have ten groups, and **percentiles** have 100 groups.

Figure 8.6.3: Menu for t quantiles, with values entered for the two-tail example.

You should confirm that what R calculates agrees with the critical values tabulated in the Table of Critical values for the t distribution provided in the Appendix.

A worked example

Let's revisit our lizard example from last time (see Chapter 8.5). Prior to conducting any inference test, we decide acceptable Type I error rates (cf. **justify alpha** discussion in Chapter 8.1); For this example, we set Type I error rate to be 1% for a 99% confidence interval.

The Rcmdr output was

```
t.test(lizz$bm, alternative='two.sided', mu=5, conf.level=.99)
data: lizz$bm
t = -1.5079, df = 7, p-value = 0.1753
alternative hypothesis: true mean is not equal to 5
99 percent confidence interval:
1.984737 6.199263
sample estimates:
mean of x
4.092
```

Sort through the output and identify what you need to know.

Question 1: What was the sample mean?

A. 5
B. -1.5079
C. 7
D. 0.1753
E. 1.984737
F. 6.199263
G. 4.092 ← Answer

Question 2: What was the most likely population mean?

A. 5 ← Answer
B. -1.5079
C. 7
D. 0.1753
E. 1.984737
F. 6.199263
G. 4.092

Question 3: This was a "one-tailed" test of the null hypothesis?

A. True
B. False ← Answer

The output states "alternative hypothesis: true mean is not equal to 5" — so it was a two-tailed test.

The 99% confidence interval $(CI_{99\%})$ is $(1.984737, 6.199263)$ which means we are 99% certain that the population mean is between 1.984737 (lower limit) and 6.199263 (the upper limit). In Chapter 8.5 we calculated the $(CI_{95\%})$ as $(2.667, 5.517)$

Confidence intervals by nonparametric bootstrap sampling

Bootstrapping is a general approach to estimation or statistical inference that utilizes random sampling with replacement (Kulesa et al. 2015). In classic frequentist approach, a sample is drawn at random from the population and assumptions about the population distribution are made in order to conduct statistical inference. By resampling with replacement from the sample many times, the bootstrap samples can be viewed as if we drew from the population many times without invoking a theoretical distribution. A clear advantage of the bootstrap is that it allows estimation of confidence intervals without assuming a particular theoretical distribution and thus avoids the burden of repeating the experiment. Which method to prefer? For cases where assumption of a particular distribution is unwarranted (e.g., what is the appropriate distribution when we compare medians among samples?), bootstrap may

be preferred (and for small data sets, percentile bootstrap may be better). We cover bootstrap sampling of confidence intervals in Chapter 19.2: Bootstrap sampling.

Conclusions

The take home message is simple.

- All estimates must be accompanied by a Confidence Interval.
- The more confident we wish to be, the wider the confidence interval will be.

Note that the confidence interval concept combines DESCRIPTION (the population mean is between these limits) and INFERENCE (and we are 95% certain about the values of these limits). It is good statistical practice to include estimates of confidence intervals for any estimate you share with readers. Any statistic that can be estimated should be accompanied by a confidence interval and, as you can imagine, formulas are available to do just this. For example, earlier this semester we calculated NNT.

Questions

1. Note in the worked example we used Type I error rate of 1%, not 5%. With a Type I error rate of 5% and sample size of 10, what will be the degrees of freedom (df) for the t distribution?
2. Considering the information in question 1, what will be the critical value of the t-distribution for
   - a one-tailed test?
   - a two-tailed test
3. To gain practice with calculations of confidence intervals, calculate the approximate confidence interval, the 95% and the 99% confidence intervals based on the t distribution, for each of the following.

- $\bar{X} = 13$, $s = 1.3$, $n = 10$
- $\bar{X} = 13$, $s = 1.3$, $n = 30$
- $\bar{X} = 13$, $s = 2.6$, $n = 10$
- $\bar{X} = 13$, $s = 2.6$, $n = 30$

4. Take at look at your answers to question 3 — what trend(s) in the confidence interval calculations do you see with respect to variability?
5. Take at look at your answers to question 3 — what trend(s) in the confidence interval calculations do you see with respect to sample size?

8.7: Chapter 8 References and Suggested Readings

Abelson, RP (1995) *Statistics As Principled Argument*. Taylor & Francis.

Allison, D. B., Cui, X., Page, G. P., Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews. Genetics* 7(1):55-65

Benjamin, D. J., et al. (2017). Redefine statistical significance. *Nature Human Behaviour*, 1.

Browner, W. S., Newman, T. B. (1987). Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *Journal of American Medical Association* 257:2459-2463.

Campbell, P., Baker, W. L., Bendel, S. D., & White, W. B. (2011). Intravenous hydralazine for blood pressure management in the hospitalized patient: its use is often unjustified. *Journal of the American Society of Hypertension*, 5(6), 473-477.

Carlson, B. (2012). Putting Oncology Patients at Risk. *Biotechnology Healthcare* 9(3): 17–21.

Cohen, J. (2016). The earth is round (p<. 05). In What if there were no significance tests? (pp. 69-82), edited by L.L. Harlow, S. A. Mulaik, J. H. Steiger. Routledge.

Cowles, M. & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist* 37:553-558.

Cox, D. R. (1982). Statistical significance testing. *British Journal of Clinical Pharmacology* 14(3):325-331.

Curran-Everett, D. (2009). Explorations in statistics: hypothesis tests and P values. *Advances in Physiology Education* 33:81-86.

DeVita VT, Chu E (2008) A history of cancer chemotherapy. *Cancer Research* 68:8643-8653

Draghici, Sorin and Khatri, Purvesh and Eklund, Aron C. and Szallasi, Zoltan (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics* 22:101-109

Flam, F. D. (2014, Sept 29). The Odds, Continually Updated. The New York Times, Retrieved from https://www.nytimes.com/2014/09/30/s...ated.html?_r=0

Fontana, L., Weiss, E. P., Villarea, D. T., Klein, S., Holloszy, J. O. (2008). Long-term effects of calorie or protein restriction on serum IGF-1 and IGFBP-3 concentration in humans. *Aging Cell* 7(5):681–687.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology* 45:135-140

Greenland, S., Poole, C. (2012). Living with P values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* 24:62-68.

Hartwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402: C47-C52.

Held, L. (2010). A nomogram for P values. *BMC Medical Research Methodology* 10:21.

Holzenberger, M. I., Dupont, J., Ducos, B., Leneuve, P., Géloën, A., Even, P. C., Cervera, P., Le Bouc, Y. (2003). IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* 421(6919):182-187.

Hubbard, R., Bayarri, M. J., Berk, K. N., Carlton, M. A. (2003). Confusion over measures of evidence (p's) versus errors (σ's) in classical statistical testing. *American Statistician* 57:171-182.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*. 2:e124. doi: 10.1371/journal.pmed.0020124

Lang, T. A., & Altman, D. G. (2015). Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *Int J Nurs Stud*, *52*(1), 5-9.

Lehmann, E. L. (1992). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two. *Journal of the American Statistical Association* 88:1242-1249.

Mukherjee, S. (2011). *The emperor of all maladies: A biography of cancer*. Scribner.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5(2):241-301.

Platt, J. R. (1964). Strong Inference. *Science* 146:347-353

Ryle, A. (2006). The relevance of evolutionary psychology for psychotherapy. *British Journal of Psychotherapy* 21(3):375-388.

Savalei, V., Dunn, E. (2015). Is the call to abandon p-values the red-herring of the replicability crisis? *Frontiers in Psychology* 6:245.

Selke, T., Bayarri, M. J., Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician* 55(2):62-71.

Simonsohn, U., Nelson, L. D., Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General* 143:534-547.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician* 73, 2019 – Issue sup1: Statistical Inference in the 21st Century.

Whitley, E, and J. Ball (2002). Statistics review 3: Hypothesis testing and P-values. *Critical Care* 6:222-225.

CHAPTER OVERVIEW

9: Categorical Data

Introduction

No doubt you have already been introduced to chi-square $\left(\chi^2\right)$ tests (click here correct pronunciation), particularly if you've had a genetics class, but perhaps you were not told why you were using the χ^2 test, as opposed to some other test, for example t-test, ANOVA, or linear regression.

Chi-square analyses are used in situations of discrete (i.e., categorical or qualitative) data types. When you can count the number of "yes" or "no" outcomes from an experiment, then you are talking about a χ^2 problem. In contrast, continuous (i.e., quantitative) data types for outcome variables would require you to use the t-test (for two groups) or the ANOVA-like procedures (for two or more groups). Chi-square tests can be applied when you have two or more treatment groups.

Two kinds of chi-square analyses

(1) We ask about the "fit" of our data against predictions from theory. This is the typical chi-square that student's have been exposed to in biology lab. If outcomes of an experiment can be measured against predictions from some theory, then this is a **goodness of fit** (gof) χ^2. Goodness of fit is introduced in Section 9.1.

(2) We ask whether the outcomes of an experiment are associated with a treatment. These are called **contingency table** problems, and they will be the subject of the next lecture. The important distinction here is that there exists no outside source of information ("theory") available to make predictions about what we would expect. Contingency tables are introduced in Section 9.2.

9.1: Chi-square test and goodness of fit

Introduction

We ask about the "fit" of our data against predictions from theory, or from rules that set our expectations for the frequency of particular outcomes drawn from outside the experiment. Three examples (**A**, **B**, and **C**) that illustrate **goodness of fit**, GOF, χ^2, follow.

A. For example, for a toss of a coin, we expect heads to show up 50% of the time. Out of 120 tosses of a fair coin, we expect 60 heads, 60 tails. Thus, our **null hypothesis** would be that heads would appear 50% of the time. If we observe 70 heads in an experiment of coin tossing, is this a significantly large enough discrepancy to reject the null hypothesis?

B. For example, simple **Mendelian genetics** makes predictions about how often we should expect particular combinations of phenotypes in the offspring when the phenotype is controlled by one gene, with 2 alleles and a particular kind of dominance.

For example, for a one-locus, two-allele system (one gene, two different copies like **R** and **r**) with complete dominance, we expect the phenotypic (what you see) ratio will be 3:1 (or $\frac{3}{4}$ round, $\frac{1}{4}$ wrinkle). Our null hypothesis would be that pea shape will obey Mendelian ratios (3:1). Mendel's round versus wrinkled peas (**RR** or **Rr** genotypes give round peas, only **rr** results in wrinkled peas).

Thus, out of 100 individuals, we would expect 75 round and 25 wrinkled. If we observe 84 round and 16 wrinkled, is this a significantly large enough discrepancy to reject the null hypothesis?

C. For yet another example, in population genetics, we can ask whether genotypic frequencies (how often a particular copy of a gene appears in a population) follow expectations from **Hardy-Weinberg** model (the null hypothesis would be that they do).

This is a common test one might perform on DNA or protein data from electrophoresis analysis. Hardy-Weinberg is a simple quadratic expansion:

If p = **allele frequency** of the first copy, and q = allele frequency of the second copy, then $p+q=1$,

Given the allele frequencies, then genotypic frequencies would be given by $1=p^2+2pq+q^2$.

Deviations from Hardy-Weinberg expectations may indicate a number of possible causes of allele change (including natural selection, genetic drift, migration).

Thus, if a gene has two alleles, a and a', with the frequency for a, $f(a)=p=0.06$ and for a', $f(a')=q=0.4$, (equivalently $q=1-p$) in the population, then we would expect 36 aa, 16 $a'a'$, and 48 aa' individuals. (Nothing changes if we represent the alleles as A and a, or some other system, e.g., dominance/recessive.)

Question. If we observe the following genotypes: 45 aa individuals, 34 aa' individuals, and 21 $a'a'$ individuals, is this a significantly large enough discrepancy to reject the null hypothesis?

Table 9.1.1. Summary of our Hardy-Weinberg question.

Genotype	Expected	Observed	O – E
aa	70	45	-25
aa'	27	34	7
$a'a'$	3	21	18
Sum	*100*	*100*	*0*

Recall from your genetics class that we can get the allele frequency values from the genotype values, e.g., $f(a)=f(aa)+\frac{1}{2}f(aa')$.

We call these chi-square tests, tests of goodness of fit. Because we have some theory, in this case Mendelian genetics, or guidance, separate from the study itself, to help us calculate expected values in a chi-square test.

Note:

The idea of fit in statistics can be reframed as how well does a particular statistical model fit the observed data. A good fit can be summarized by accounting for the differences between the observed values and the comparable values predicted by the model.

χ^2 goodness of fit

For k groups, the equation for the chi-square test may be written as

$$\chi^2=\sum_{i=1}^{k}\frac{\left(f_i-\hat{f}_i\right)^2}{\hat{f}_i}$$

where f_i is the frequency (count) observed (in class i) and $\hat{f}_i$ is the frequency (count) expected if the null hypothesis is true, summed over all k groups. Alternatively, here is a format for the same equation that may be more familiar to you… ?

$$\chi^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$$

where O_i is the frequency (count) observed (in class i) and E_i is the frequency (count) expected if the null hypothesis is true.

The degrees of freedom, df, for the GOF χ^2 are simply the number of categories minus one, $k-1$.

Explaining GOF

Why am I using the phrase "goodness of fit?" This concept has broad use in statistics, but in general it applies when we ask how well a statistical model fits the observed data. The chi-square test is a good example of such tests, and we will encounter other examples too. Another common goodness of fit is the coefficient of determination, which will be introduced in linear regression sections. Still other examples are the **likelihood ratio test**, **Akaike Information Criterion (AIC)**, and **Bayesian Information Criterion (BIC)**, which are all used to assess fit of models to data. (See Graffelman and Weir [2018] for how to use AIC in the context of testing for Hardy Weinberg equilibrium.) At least for the chi-square test it is simple to see how the test statistic increases from zero as the agreement between observed data and expected data depart, where zero would be the case in which all observed values for the categories exactly match the expected values.

This test is designed to evaluate whether or not your data agree with a theoretical expectation (there are additional ways to think about this test, but this is a good place to start). Let's take our time here and work with an example. The other type of chi-square problem or experiment is one for the many types of experiments in which the response variable is discrete, just like in the GOF case, but we have no theory to guide us in deciding how to obtain the expected values. We can use the data themselves to calculate expected values, and we say that the test is "contingent" upon the data, hence these types of chi-square tests are called **contingency tables**.

You may be a little concerned at this point that there are two kinds of chi-square problems, goodness of fit and contingency tables. We'll deal directly with contingency tables in the next section, but for now, I wanted to make a few generalizations.

1. Both goodness of fit and contingency tables use the same chi-square equation and analysis. They differ in how the degrees of freedom are calculated.
2. Thus, what all chi-square problems have in common, whether goodness of fit or contingency table problems, are:
    1. You must identify what types of data are appropriate for this statistical procedure? Categorical (**nominal data type**).
    2. As always, a clear description of the hypotheses being examined.

For goodness of fit chi-square test, the most important type of hypothesis is called a Null Hypothesis: In most cases the Null Hypothesis (H_O) is "no difference" "no effect"…. If H_O is concluded to be false (rejected), then an alternate hypothesis (H_A) will be assumed to be true. Both are specified before tests are conducted. All possible outcomes are accounted for by the two hypotheses.

From above, we have

A. H_O: Fifty out of 100 tosses will result in heads.

H_A: Heads will not appear 50 times out of 100.

B. H_O: Pea shape will equal Mendelian ratios (3:1).

H_A: Pea shape will not equal Mendelian ratios (3:1).

C. H_O: Genotypic frequencies will equal Hardy-Weinberg expectations.

H_A: Genotypic frequencies will not equal Hardy-Weinberg expectations

Assumptions: In order to use the chi-square, there must be two or more categories. Each observation must be in one and only one category. If some of the observations are truly halfway between two categories then you must make a new category (e.g. low, middle, high) or use another statistical procedure. Additionally, your expected values are required to be integers, not ratio. The number of observed and the number of expected must sum to the same total.

How well does data fit the prediction?

Frequentist approach interprets the test as, how well does the data fit the null hypothesis, $P\left(data|H_O\right)$? When you compare data against a theoretical distribution (e.g., Mendel's hypothesis predicts the distribution of progeny phenotypes for a particular genetic system), you test the fit of the data against the model's predictions (expectations). Recall that the Bayesian approach asks how well does the model fit the data?

A. 120 tosses of a coin, we count heads 70/120 tosses.

	Expected	Observed
Heads	60	70
Tails	60	50
n	*120*	*120*

$$\chi^2 = \frac{(70-60)^2}{60} + \frac{(50-60)^2}{60} = 1.667 + 1.667 = 3.333$$

B. A possible Mendelian system of inheritance for a one-gene, two-allele system with complete dominance, observe the phenotypes.

	Expected	Observed
Round	75	84
Wrinkled	25	16
n	*100*	*100*

$$\chi^2 = \frac{(84-75)^2}{75} + \frac{(16-25)^2}{25} = 1.080 + 3.240 = 4.320$$

C. A possible Mendelian system of inheritance for a one-gene, two-allele system with complete dominance, observe the phenotypes.

	Expected	Observed
p^2	70	45
$2pq$	27	34
q^2	3	21
n	*100*	*100*

$$\chi^2 = \frac{(45-70)^2}{70} + \frac{(34-27)^2}{27} + \frac{(21-3)^2}{3} = 8.93 + 1.82 + 108 = 118.74$$

For completeness, instead of a goodness of fit test we can treat this problem as a test of independence, a contingency table problem. We'll discuss contingency tables more in the next section, but or now, we can rearrange our table of observed genotypes for problem C, as a 2 × 2 table:

	Maternal a'	**Paternal a'**
Maternal a	45	17
Paternal a	17	21

The contingency table is calculated the same way as the GOF version, but the degrees of freedom are calculated differently: df = number of rows – 1 multiplied by the number of columns – 1.

$$df = (rows - 1)(columns - 1)$$

Thus, for a 2 × 2 table the df are always equal to 1.

Note that the chi-square value itself says nothing about how any discrepancy between expectation and observed genotype frequencies come about. Therefore, one can rearrange the χ^2 equation to make clear where deviance from equilibrium, D, occurs for the heterozygote (het). We have

$$\chi^2 = \frac{D^2}{p^2 q^2 n}$$

where D^2 is equal to $D^2 = \frac{1}{2}(O_{het} - E_{het})$.

Carry out the test and interpret results

What was just calculated? The chi-square, χ^2, **test statistic**.

Just like t-tests, we now want to compare our test statistic against a **critical value** — calculate degrees of freedom ($df = k - 1$ (k equals the numbers of categories), and set a rejection level, **Type I error rate**. We typically set the Type I error rate at 5%. A table of critical values for the chi-square test is available in Appendix: Table of Chi-square critical values.

Obtaining Probability Values for the χ^2 goodness-of-fit test of the null hypothesis:

As you can see from the equation of the chi-square, a perfect fit between the observed and the expected would be a chi-square of zero. Thus, asking about statistical significance in the chi-square test is the same as asking if your test statistic is significantly greater than zero.

The chi-square distribution is used and the critical values depend on the degrees of freedom. Fortunately, for χ^2 and other statistical procedures we have tables that will tell us what the probability is of obtaining our results when the null hypothesis is true (in the population).

Here is a portion of the chi-square critical values for probability that your chi-square test statistic is less than the critical value.

Figure 9.1.1: A portion of the chi-square critical values table.

For the first example (**A**), we have $df = 2 - 1 = 1$ and we look up the critical value corresponding to the probability in which Type I = 5% are likely to be smaller iff ("if and only if") the null hypothesis is true. That value is 3.841; our test statistic was 3.330, and therefore smaller than the critical value, so we do not reject the null hypothesis.

Interpolating p-values

How likely is our test statistic value of 3.333 and the null hypothesis was true? (Remember, "true" in this case is a shorthand for our data was sampled from a population in which the HW expectations hold). When I check the table of critical values of the chi-square test for the "exact" p-value, I find that our test statistic value falls between a p-value 0.10 and 0.05 (represented in the table below). We can **interpolate**

> **Note:**
>
> Interpolation refers to any method used to estimate a new value from a set of known values. Thus, interpolated values fall between known values. **Extrapolation** on the other hand refers to methods to estimate new values by extending from a known sequence of values.

statistic	p-value
3.841	0.05
3.333	x
2.706	0.10

If we assume the change in probability between 2.706 and 3.841 for the chi-square distribution is linear (it's not, but it's close), then we can do so simple interpolation.

We set up what we know on the right hand side equal to what we don't know on the left hand side of the equation,

$$\frac{x - 0.10}{0.05 - 0.10} = \frac{3.333 - 2.706}{3.841 - 2.706}$$

and solve for x. Then, x is equal to 0.0724.

R function `pchisq()` gives a value of `p = 0.0679` . Close, but not the same. Of course, you should go with the result from R over interpolation; we mention how to get the approximate p-value by interpolation for completeness, and, in some rare instances, you might need to make the calculation. Interpolating is also a skill used to provide estimates where the researcher needs to estimate (impute) a missing value.

Interpreting p-values

What does it mean to reject the null hypothesis? These types of tests are called goodness of fit in this sense — if your data agree with the theoretical distribution, then the difference between observed and expected should be very close to zero. If it is exactly zero, then you have a perfect fit. In this case, then we say that the ratio of `heads:tails` do not differ significantly from the `50:50` expectation if we accept the null hypothesis.

You should try the other examples yourself! As a hint, the degrees of freedom are 1 for example B and 2 for example C.

R code

Printed tables of the critical values from the chi-square distribution, or for any statistical test for that matter are fine, but with your statistical package R and `Rcmdr` , you have access to the critical value and the p-value of your test statistic simply by asking. Here's how to get both.

First, let's get the critical value.

Rcmdr: Distributions → Continuous distributions → Chi-squared distribution → Chi-squared quantiles

Figure 9.1.2: R Commander menu for Chi-squared quantiles.

I entered "0.05" for the probability because that's my Type I error rate α. Enter "1" for Degrees of freedom, then click "upper tail" because we are interested in obtaining the critical value for α. Here's R's response when I clicked "OK."

```
qchisq(c(0.05), df=1, lower.tail=FALSE)
[1] 3.841459
```

Next, let's get the exact P-value of our test statistic. We had three from three different tests: $\chi^2 = 3.333$ for the coin-tossing example, $\chi^2 = 4.320$ for the pea example, and $\chi^2 = 7.8955$ for the Hardy-Weinberg example.

Rcmdr: Distributions → Continuous distributions → Chi-squared distribution → Chi-squared probabilities…

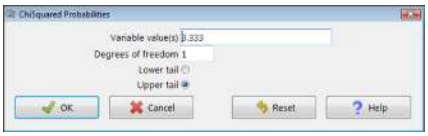


Figure 9.1.3: R Commander menu for Chi-squared probabilities.

I entered "3.333" because that is one of the test statistics I want to calculate for probability and "1" for Degrees of freedom because I had $df = k - 1 = 1$ for this problem. Here's R's response when I clicked "OK."

```
pchisq(c(3.333), df=1, lower.tail=FALSE)
[1] 0.06790291
```

I repeated this exercise for $\chi^2 = 4.320$. I got $p = 0.03766692$ for $\chi^2 = 7.8955$ I got $p = 0.004955794$

How to get the goodness of fit χ^2 in Rcmdr.

R provides the goodness of fit χ^2 (the command is `chisq.text()`), but Rcmdr thus far does not provide a menu option to link to the function. Instead, R Commander provides a menu for contingency tables, which also is a chi-square test, but is used where no theory is available to calculate the expected values (see Chapter 9.2). Thus, for the goodness of fit chi-square, we will need to by-pass Rcmdr in favor of the script window. Honestly, other options are as quick or quicker: calculate by hand, use a different software (e.g, Microsoft Excel), or many online sites provide JavaScript tools (e.g., www.graphpad.com).

So how to get the goodness of fit chi-square while in R? Here's one way. At the command line, type

```
chisq.test (c(O1, O2, ... On), p = c(E1, E2, ... En))
```

where O1, O2, … On are observed counts for category 1, category 2, up to category n, and E1, E2, … En are the expected proportions for each category. For example, consider our Heads/Tails example above (problem A).

In R, we write and submit

```
chisq.test(c(70,30),p=c(1/2,1/2))
```

R returns

```
chisq.test(c(70,30),p=c(1/2,1/2))
Chi-squared test for given probabilities.
data: c(70, 30)
X-squared = 16, df = 1, p-value = 0.00006334
```

Easy enough. But not much detail — details are available with some additions to the R script. I'll just link you to a nice website that shows how to add to the output so that it looks like the one below.

```
mike.chi <- chisq.test(c(70,30),p=c(1/2,1/2))
```

Let's explore one at time the contents of the results from the chi square function.

```
names(mike.chi)   #The names function
[1] "statistic" "parameter" "p.value" "method" "data.name" "observed"
[7] "expected" "residuals" "stdres"
```

Now, call each name in turn.

```
mike.chi$residuals
[1] 2.828427 -2.828427
mike.chi$obs
[1] 70 30
mike.chi$exp
[1] 50 50
```

Note:

"**residuals**" here simply refers to the difference between observed and expected values. Residuals are an important concept in regression, see Ch. 17.5 – Testing regression coefficients

And finally, let's get the summary output of our statistical test.

```
mike.chi
Chi-squared test for given probabilities.
data: c(70, 30)
X-squared = 16, df = 1, p-value = 6.334e-05
```

χ^2 GOF and spreadsheet apps

Easy enough with R, but it may even easier with other tools. I'll show you how to do this with spreadsheet apps and with and online at graphpad.com.

Let's take the pea example above. We had 16 wrinkled, 84 round. We expect 25% wrinkled, 75% round.

Now, with R, we would enter

```
chisq.test(c(16,80),p=c(1/4,3/4))
```

and the R output would be

```
Chi-squared test for given probabilities
data: c(16, 80)
X-squared = 3.5556, df = 1, p-value = 0.05935
```

Microsoft Excel and the other spreadsheet programs (Apple Numbers, Google Sheets, LibreOffice Calc) can calculate the goodness of fit directly; they return a P-value only. If the observed data are in cells A1 and A2, and the expected values are in B1 and B2, then use the procedure `=CHITEST(A1:A2,B1:B2)` .

	A	B	C	D
1	80	75		
2	16	25		=CHITEST(A1:A2,B1:B2)

The P-value (but not the Chi-square test statistic) is returned. Here's the output from Calc.

	A	B	C	D
1	80	75		
2	16	25		0.058714340077662

You can get the critical value from MS Excel `(=CHIINV(alpha, df)` , returns the critical value), and the exact probability for the test statistic `=CHIDIST(x,df)` , where *x* is your test statistic. Putting it all together, here is what a general spreadsheet template for χ^2 goodness of fit calculations calculations of test statistic and p-value might look like:

	A	B	C	D	E
1	**f1**	0.75			
2	**f2**	0.25			
3	**N**	=SUM(A5,A6)			
4	**Obs**	**Exp**	**Chi.value**	**Chi.sqr**	
5	80	=B1*B3	=((A5-B5)^2)/B5	=SUM(C5,C6)	
6	16	=B2*B3	=((A6-B6)^2)/B6		=CHIDIST(D5,COUNT(A5:A6-1)
7					
8					

Microsoft Excel can be improved by writing macros, or by including available add-in programs, such as the free PopTools, which is available for Microsoft Windows 32-bit operating systems only.

Another option is to take advantage of the internet — again, many folks have provided java or JavaScript-based statistical routines for educational purposes. Here's an easy one to use www.graphpad.com.

In most cases, I find the chi-square goodness-of-fit is so simple to calculate by hand that the computer is redundant.

Questions

1. A variety of p-values were reported on this page with no attempt to reflect significant figures or numbers of digits (see Chapter 8.2). Provide proper significant figures and numbers of digits as if these p-values were reported in a science journal.

a. 0.0724
b. 0.0679
c. 0.03766692
d. 0.004955794
e. 0.00006334

f. 6.334e-05
g. 0.05935
h. 0.058714340077662

2. For a mini bag of M&M candies, you count 4 blue, 2 brown, 1 green, 3 orange, 4 red, and 2 yellow candies.

a. What are the expected values for each color?
b. Calculate χ^2 using your favorite spreadsheet app (e.g., Numbers, Excel, Google Sheets, LibreOffice Calc)
c. Calculate χ^2 using R (note R will reply with a warning message that the "Chi-squared approximation may be incorrect"; see 9.3: Yates continuity correction)
d. Calculate χ^2 using Quickcalcs at graphpad.com
e. Construct a table and compare p-values obtained from the different applications

3. CYP1A2 is an enzyme involved with metabolism of caffeine. Folks with C at SNP rs762551 have higher enzyme activity than folks with A. Populations differ for the frequency of C. Using R or your favorite spreadsheet application, compare the following populations against global frequency of C that is 33% (frequency of A is 67%).

a. 286 persons from Northern Sweden: f(C) = 26%, f(A) = 73%
b. 4532 Native Hawaiian persons: f(C) = 22%, f(A) = 78%
c. 1260 Native American persons: f(C) = 30%, f(A) = 70%
d. 8316 Native American persons: f(C) = 36%, f(A) = 64%
e. Construct a table and compare p-values obtained for the four populations.

9.2: Chi-square contingency tables

Introduction

We just completed a discussion about goodness of fit tests, inferences on categorical traits for which a theoretical distribution or expectation is available. We offered Mendelian ratios as an example in which theory provides clear-cut expectations for the distribution of phenotypes in the F_2 offspring generation. This is an **extrinsic model** — theory external to the study guides in the calculation of expected values — and it reflects a common analytical task in epidemiology.

But for many other kinds of categorical outcomes, no theory is available. Today, most of us accept the link between tobacco cigarette smoking and lung cancer. The evidence for this link was reported in many studies, some better designed than others.

Consider the famous Doll and Hill (1950) report on smoking and lung cancer. They reported (Table IV Doll and Hill 1950), for men with lung cancer, only two out of 649 were non-smokers. In comparison, 27 case-control patients (i.e., patients in the hospital but with other ailments, not lung cancer) were nonsmokers, but 622 were smokers (Table 9.2.1).

Table 9.2.1. Smoking and lung cancer (Doll and Hill, 1950).

	Smokers	Non-smokers
Lung cancer	647	2
Case controls, no lung cancer	622	27

A more recent example from the study of the efficacy of St John's Wort (*Hypericum perforatum*) as a treatment for major depression (Shelton et al 2001, Apaydin et al. 2016). Of patients who received St. John's Wort over 8 weeks, 14 were deemed improved while 98 did not improve. In contrast 5 patients who received the placebo were deemed improved while 102 did not improve (Table 9.2.2).

Table 9.2.2. St. John's Wort and depression.

	Improved	Not improved
St. John's Wort	14	98
Placebo	5	102

Note:

The St. John's Wort problem is precisely a good time to remind any reader of Mike's Biostatistics Book: under no circumstances is medical advice implied from my presentation. From the National Center for Complementary and Integrative Medicine: "It has been clearly shown that St. John's wort can interact in dangerous, sometimes life-threatening ways with a variety of medicines."

These kinds of problems are direct extensions of our risk analysis work in Chapter 7. Now, instead of simply describing the differences between case and control groups by Relative Risk Reduction or odds ratios, we instruct how to do inference on the risk analysis problems.

Thus, faced with an absence of coherent theory as guide, the data themselves can be used (**intrinsic model**), and at least for now, we can employ the χ^2 test.

Note:

The preferred analysis is to use a logistic regression approach, Chapter 18.3, because additional covariates can be included in the model.

In this lesson we will learn how to extend the analyses of categorical data to cases where we do not have a prior expectation — we use the data to generate the tests of hypotheses. This would be an intrinsic model. One of the most common two-variable categorical analysis involves 2×2 contingency tables. We may have one variable that is the treatment variable and the second variable is the outcome variable. Some examples are provided in Table 9.2.3.

Table 9.2.3. Some examples of treatment and outcome variables suitable for contingency table analysis.

Treatment Variable and Levels	Outcome Variable(s)	Reference
Lead exposure Levels: Low, Medium, High	Intelligence and development	Bellinger et al 1987
Wood preservatives Levels: Borates vs. Chromated Copper Arsenate	Effectiveness as fungicide	Hrastnik et al 2013
Antidepressants Levels: St. John's Wort, conventional antidepressants, placebo	Depression relief	Linde et al 2008
Coral reefs Levels: Protected vs Unprotected	Fish community structure	Guidetti 2006
Aspirin therapy Levels: low dose vs none	Cancer incidence in women	Cook et al 2013
Aspirin therapy Levels: low dose vs none	Cancer mortality in women	Cook et al 2013

Table 9.2.3 holds examples of published studies returned from a quick PubMed search; there are many examples (meta-analysis opportunities!). However, while they all can be analyzed by contingency tables analysis, they are not exactly the same. In some cases, the treatments are **fixed effects**, where the researcher selects the levels of the treatments. In other cases, each of these treatment variables need not be actual treatments (in the sense that an experiment was conducted), but it may be easier to think about these as types of experiments. These types of experiments can be distinguished by how the sampling from the reference population were conducted. Before we move on to our main purpose, to discuss how to calculate contingency tables, I wanted to provide some experimental design context by introducing two kinds of sampling (Chapter 5.5).

Our first kind of sampling scheme is called **unrestricted sampling**. In unrestricted sampling, you collect as many subjects (observations) as possible, then assign subjects to groups. A common approach would be to sample with a grand total in mind; for example, your grant is limited and so you only have enough money to make 1000 copies of your survey and you therefore approach 1000 people. If you categorize the subjects into just two categories (e.g., Liberal, Conservative), then you have a binomial sample. If instead you classify the subjects according to a number of variables, e.g., Liberal or Conservative, Education levels, income levels, home owners or renters, etc., then this approach is called multinomial sampling. The point in either case is that you have just utilized a multinomial sampling approach. The aim is to classify your subjects to their appropriate groups after you have collected the sample.

Logically, what must follow unrestricted sampling would be **restricted sampling**. Sampling would be conducted with one set of "**marginal totals**" fixed. Margins refers to either the row totals or to the column totals. The sampling scheme is referred to as compound multinomial. The important distinction from the other two types is that the number of individuals in one of the categories is fixed ahead of time. For example, in order to determine if smoking influences respiratory health, you approach as many people as possible to obtain 100 smokers.

The contingency table analysis

We introduced the 2×2 contingency table (Table 9.2.4) in Chapter 7.

Table 9.2.4. 2×2 contingency table.

	Outcome		
Exposure or Treatment group	**Yes**	**No**	**Marginal total**
Exposure (Treatment)	**a**	**b**	Row1 = **a** + **b**
Nonexposure (Control)	**c**	**d**	Row2 = **c** + **d**
Marginal total	Column1 = **a** + **c**	Column2 = **a** + **c**	**N**

Contingency tables are all of the form (Table 9.2.5)

Table 9.2.5. Basic format of a 2×2 table

	Outcome 1	**Outcome 2**
Treatment 1	Yes	No
Treatment 2	Yes	No.

Regardless of how the sampling occurred, the analysis of the contingency table is the same; mechanically, we have a chi-square type of problem. In both contingency table and the "goodness of fit" Chi-Square analyses the data types are discrete categories. The difference between GOF and contingency table problems is that we have no theory or external model to tell us about what to expect. Thus, we calculate expected values and the degrees of freedom differently in contingency table problems. To learn about how to perform contingency tables we will work through an example, first the long way, and then using a formula and the **a, b, c,** and **d** 2×2 table format. Of course, R can easily do contingency tables calculations for us.

Doctors noticed that a new form of Hepatitis (HG, now referred to as GB virus C), was common in some HIV+ populations (Xiang et al 2001). Accompanying the co-infection, doctors also observed an inverse relationship between HG loads and HIV viral loads: patients with high HG titers often had low HIV levels. Thus, the question was whether co-infection alters the outcome of patients with HIV — do HIV patients co-infected with this HG progress to AIDS and mortality at rates different from non-infected HIV patients? I've represented the Xiang et al (2001) data in Table 9.2.6.

Table 9.2.6. Progression of AIDS for patients co-infected with HG [GB virus C], Xiang et al data (2001).

	Lived	**Died**	**Row totals**
HG+	103	41	144
HG-	95	123	218
Column totals	198	164	362

Note:

HG virus is no longer called a hepatitis virus, but instead is referred to as GB virus C, which, like hepatitis viruses, is in the *Flaviviridae* virus family. For more about the GB virus C, see the review article by Bhattarai and Stapleton (2012).

Our question was: Do HIV patients co-infected with this HG progress to AIDS and mortality at rates different from non-infected HIV patients? A typical approach to analyze such data sets is to view the data as discrete categories and analyze with a contingency table. So we proceed.

Set up the table for analysis

Rules of contingency tables (see Kroonenberg and Verbeek 2018).

What follows is a detailed walk though setting up and interpreting a 2×2 table. Note that I maintain our a, b, c, d cell order, with row 1 referencing subjects exposed or part of treatment group and row 2 referencing subjects not exposed (or part of the control group), as introduced in Chapter 7.

The Hepatitis G data from Xiang et al 2001 data, arranged in 2×2 format (Table 9.2.7).

Table 9.2.7. Format of 2×2 table Xiang et al (2001) dataset.

	Lived	**Died**	**Row totals**
HG+	**a**	**b**	a + b
HG-	**c**	**d**	c + d
Column totals	a + c	b + d	N

The data are placed into the cells labeled **a**, **b**, **c**, and **d**.

Cell **a** : The number of HIV+ individuals infected with HG that lived beyond the end of the study.

Cell **b** : The number of HIV+ individuals infected with HG that died during the study.

Cell **c** : The number of HIV+ individuals not infected with HG that lived.

Cell **d** : The number of HIV+ individuals not infected with HG that died.

This is a contingency table because the probability of living or dying for these patients may have been contingent on coinfection with hepatitis G.

State the examined hypotheses

H_O: There is NO association between the probability of living and coinfection with hepatitis G.

For a contingency table this means that there should be the same proportion of individuals with and without hepatitis G that either lived or died (1:1:1:1).

Now, compute the Expected Values in the Four cells (a, b, c, d)

1. Calculate the Expected Proportion of Individuals that lived in Entire Sample:
   Column 1 Total / Total = Expected Proportion of those that lived beyond the study.
2. Calculate Expected Proportion of Individuals that died in Entire Sample:
   Column 2 Total / Total = Expected Proportion that died during the study.
3. Calculate the Expected Proportion of Individuals that lived and were HG+
   Expected Proportion of living (step 1 above) × Row 1 Total = Expected For Cell A.
4. Calculate the Expected Proportion of Individuals that died and were HG+
   Expected Proportion died × Row 1 Total = Expected For Cell B
5. Calculate the Expected Proportion of surviving individuals that were HG-
   Expected Proportion lived × Row 2 Total = Expected For Cell C
6. Calculate the Expected Proportion of individuals that died that were HG-
   Expected Proportion that died × Row 2 Total = Expected For Cell D

Yes, this can get a bit repetitive! But, we are now done — remember, we're working through this to review how the intrinsic model is applied to obtain expected values.

Summary of what we've done so far

We now have arranged our observations in a 2×2 table, and calculated the expected proportions, i.e., the Expected values under the Null Hypothesis.

Now, we proceed to conduct the chi-square test of the null hypothesis — The observed values may differ from these expected values, meaning that the Null Hypothesis is False.

Table 9.2.8. Copy of Table 9.2.6 dataset.

	Lived	Died	Row totals
HG+	103	41	144
HG-	95	123	218
Column totals	198	164	362

Get the Expected Values

1. Calculate the Expected Proportion of Individuals that lived in Entire Sample:
   Column 1 Total (198) / Total (362) = 0.5470
2. Calculate Expected Proportion of Individuals that died in Entire Sample:
   Column 2 Total (164) / Total (362) = 0.4530
3. Calculate the Expected Proportion of Individuals that lived and were HG+
   Expected Proportion of living (0.547) × Row 1 Total (218) = Cell A = 119.25
4. Calculate the Expected Proportion of Individuals that died and were HG+
   Expected Proportion died (0.453) × Row 1 Total (218) = Cell B = 98.75.
5. Calculate the Expected Proportion of surviving individuals that were HG-
   Expected Proportion lived (0.547) × Row 2 Total (144) = Cell C = 78.768
6. Calculate the Expected Proportion of individuals that died that were HG-
   Expected Proportion that died (0.453) × Row 2 Total = Cell D = 65.232

Thus, we have Table 9.2.9.

Table 9.2.9. Expected values for Xiang et al (2001) data set (Table 6 and Table 8).

	Lived	Died	Row totals
HG+	78.768	65.232	144
HG-	119.25	98.75	218
Column totals	198	164	362

Now we are ready to calculate the Chi-Square Value

Recall the formula for the chi-square test:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Table 9.2.10. Worked contingency table for Xiang et al (2001) data set (Table 6 and Table 8).

Cell	$(O_i - E_i)^2$	χ^2
a	$(103 - 87.768)^2 = 78.768$	7.4547
b	$(41 - 65.232)^2 = 65.232$	9.002
c	$(95 - 119.25)^2 = 119.25$	4.93134
d	$(123 - 98.75)^2 = 98.75$	5.95506
	$\chi^2 =$	**27.3452**

Adding all the parts we have the chi-square test statistic $\chi^2 = 27.3452$. To proceed with the inference, we test using the chi-square distribution.

Determine the Critical Value of χ^2 test and evaluate the null hypothesis

Now recall that we can get the critical value of this test in one of two (related!) ways. One, we could run this through our statistical software and get the p-value, the probability of our result and the null hypothesis is true. Second, we look up the critical value from an appropriate statistical table. Here, we present option 2.

We need the Type I error rate α and calculate the degrees of freedom for the problem. By convention, we set $\alpha = 0.05$. Degrees of freedom for contingency table is calculated as

Degrees of Freedom = $(\#\text{ rows} - 1) \times (\#\text{ columns} - 1)$

and for our example → $DF = (2-1)\times(2-1) = 1$.

Note. How did we get the 1 degree of freedom? Aren't there 4 categories and shouldn't we therefore have $df = k-1 = 3$? We do have four categories, yes, but the categories are not independent. We need to take this into account when we evaluate the test, and this lack of independence is accounted for by the loss of degrees of freedom.

What exactly is not independent about our study?

1. The total number of observations is set in advance from the data collection.
2. We also have the column and row totals set in advance. The number of HIV individuals with or without Hepatitis G infection was determined at the beginning of the experiment. The number of individuals that lived or died was also determined before we we conducted the test.
3. Then the first cell (let us make that cell A) can still vary any where from 0 to N_i (sample size of the first drug).
4. Once the first cell (A) is determined then the next cell in that row (B) will have to add up to N_i.
5. Also the other cell in the same column as the first cell (C) must add up to the Column 1 Total.
6. Lastly, the last cell (D) must have the Row 1 Total add up to the correct number.

All this translates into there being only one cell that is FREE TO VARY, hence, only one degree of freedom.

Get the Critical Value from a chi-square distribution table: For our example, look up the critical value for $DF = 1$, $\alpha = 0.05$, you should get 3.841. The 3.841 is the value of the chi-square we would expect at 5% and the null hypothesis is in fact true condition in the population.

Once we obtain the critical value we simply use the previous statement regarding the probability of the Null Hypothesis being TRUE. We reject the Null Hypothesis when the calculated χ^2 test statistic is greater than the Critical Value. We Accept the Null Hypothesis when the calculated χ^2 is less than the Critical Value. In our example we got 27.3452 for our calculated test statistic; thus we reject the null hypothesis and conclude that, at least for the sample in this study, there was an association between the probability of patients living past the study period and the presence of hepatitis G.

χ^2 from **a**, **b**, **c**, and **d** 2×2 table format formula

If a hand calculation is required, a simpler formula to use is \[\chi^{2} = \frac{(ad - bc)^{2} \cdot N}{r_{1} \cdot r_{2} \cdot c_{1} \cdot c_{2} \nonumber\]

where N is the sum of all observations in the table, r_1 and r_2 are the marginal row totals, and c_1 and c_2 are the marginal column totals from the 2×2 contingency table (e.g., Table 9.2.4). If you are tempted to try this in a spreadsheet, and assuming your entries for a, b, c, d, and N look like Table 9.2.11, then a straightforward interpretation requires referencing five spreadsheet cells no less than 13 times! Not to mention a separate call to the χ^2 distribution to calculate the p-value (one-tailed test).

> **Note:**
>
> This formulation was provided as equation 1 in Yates 1984 (referred to in Serra 2018), but is likely found in the earlier papers on χ^2 dating back to Pearson.

Table 9.2.11. Example spreadsheet with formulas for odds ratio (OR), Pearson's χ^2, and p-value from χ^2 distribution.

	A	B	C	D	E	F	G
1							
2	a	103			OR	=(B2*B5)/(B3*B4)	
3	b	41					
4	c	95					
5	d	123			chisq, p	=(((B2*B5-B3*B4)^2)*B6)/ (SUM(B2:B3)*SUM(B4:B5)* SUM(B2,B4)*SUM(B3,B5))	=CHIDIST(F5,1)
6	N	632					

Now, let's see how easy this is to do in R.

χ^2 effect size

Inference on hypotheses about association, there's statistical significance — evaluate p-value compared to Type I error, e.g., 5% — and then there's biological importance. The concept of effect size is an attempt to communicate the likely importance of a result. Several statistics are available to communicate effect size: for χ^2 that's ϕ, phi.

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

where N is the total. Phi for our example is

$$\phi = \sqrt{\frac{27.3452}{362}}$$

R code:

```
sqrt(27.3452/362)
```

returns

```
[1] "1.7e-07"
```

Effect size statistics typically range from 0 to 1; Cohen (1992) suggested the following interpretation:

Effect size	Interpretation
< 0.2	Small, weak effect
0.5	Moderate effect
> 0.8	Large effect

For our example, ϕ, the effect size of the association between co-infection with GB virus C and mortality, was weak in the patients with HIV infection.

Contingency table analyses in Rcmdr

Assuming you have already summarized the data, you can enter the data directly in the `Rcmdr` contingency table form. If your data are not summarized, then you would use `Rcmdr's` Two-way table… for this. We will proceed under the assumption that you have already obtained the frequencies for each cell in the table.

Rcmdr: Statistics → Contingency tables → Enter and analyze two-way table…

Here you can tell R how many rows and how many columns.

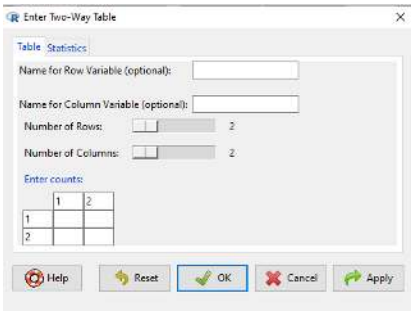


Figure 9.2.1: Screenshot R Commander menu for 2×2 data entry with counts.

The default is a 2×2 table. For larger tables, use the sliders to increase the number of rows, columns, or both.

Next, enter the counts. You can edit every cell in this table, including the headers. In the next panel, I will show you the data entry and options. The actual calculation of the χ^2 test statistic is done when you select the "Chi-square test of independence" check-box.

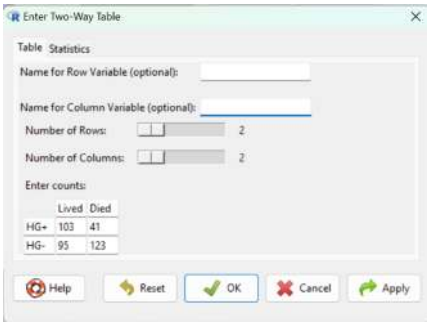


Figure 9.2.2: Display of Xiang et al data entered into R Commander menu.

After entering the data, click on the Statistics tab (Fig. 3).

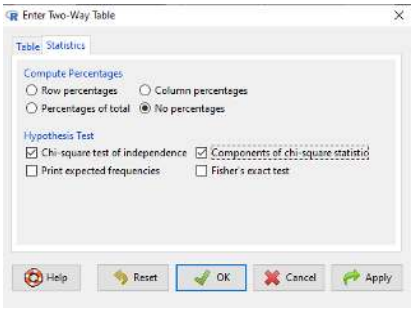


Figure 9.2.3: Screenshot Statistics options for contingency table.

If you also select "Components of chi-square statistic" option, then R will show you the contributions of each cell $(O-E)$ towards the chi-square test statistic value. This is helpful to determine if rejection of the null is due to a subset of the categories, and it also forms the basis of the heterogeneity tests, a subject we will pick up in the next section.

Here's the R output from R Commander. Note the " `2, 2, byrow=TRUE` " instructions (check out R help to confirm what these settings confirm).

```
> .Table <- matrix(c(103,41,95,123), 2, 2, byrow=TRUE)
> dimnames(.Table) <- list("rows"=c("HG+", "HG-"), "columns"=c("Lived", "Died"))

> .Table # Counts
     columns
rows Lived Died
 HG+   103   41
 HG-    95  123

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test

Pearson's Chi-squared test
```

```
data: .Table
X-squared = 27.339, df = 1, p-value = 0.0000001708

> round(.Test$residuals^2, 2) # Chi-square Components
    columns
rows Lived Died
HG+   7.46 9.00
HG-   4.93 5.95

> remove(.Test)
> remove(.Table)
```

end R output.

Note:

R Commander often cleans up after itself by removing objects like .Test and .Table. This is not necessary for the code to work, but does make it easier for you to go back and modify the code without worrying about confusing objects.

There is lots of output, but take a deep breath and remember… The minimum output we need to look at is…?

Value of the test statistic	27.339
Degrees of freedom	1
p-value	0.0000001708

We can see that the p-value, 1.7^{-7}, is much less than Type 1 error $\alpha = 0.05$; thus, by our decision criterion we reject the null hypothesis (provisionally of course, as science goes).

Note:

It's simple enough to get R to report numbers as you need. For example, R code

```
myNumber = 0.0000001708
format(myNumber, scientific = TRUE, digits=2)
```

returns

```
[1] "1.7e-07"
```

Questions

1. Instead of R Commander, try the contingency table problem in R directly.

   ```
   myData <- matrix(c(103,41,95,123))
   chisq.test(myData)
   ```

   Is this the correct χ^2 contingency table analysis? Why or why not?

2. For many years National Football League games that ended in ties went to "sudden death," where the winner was determined by the first score in the extra period of play, regardless of whether or not the other team got an opportunity to possess the ball on offense. Thus, in more than 100 games (140), the team that won the coin toss and therefore got the ball first in overtime won the game either following a kicked field goal or after a touchdown was scored. In 337 other games under this system, the outcome was not determined by who got the ball first. Many complained that the "sudden death" format was unfair and in 2013 the NFL changed its overtime rules. Beginning 2014 season, both teams got a chance to possess the ball in overtime, unless the team that won the coin toss also went on to score a touchdown, at which time that team would be declared the winner. In this new era of overtime rules ten teams that won the coin toss went on to score a touchdown in their first possession and therefore win the game, whereas in 54 other overtime games, the outcome was decided after both teams had a chance on offense (data as of 1 December 2015). These data may be summarized in the table:

	First possession win?	
	Yes	No
Coin flip years	140	337
New era	10	54

   A. What is the null hypothesis?
   B. Which is more appropriate: to calculate an odds ratio or to calculate an RRR?
   C. This is a contingency table problem. Explain why
   D. Conduct the test of the null hypothesis.
   E. What is the value of the test statistic? Degrees of freedom? P-value?
   F. Evaluate the results of your analysis — do you accept or reject the null hypothesis?

3. Return to the Doll and Hill (1950) data: 2 men with lung cancer were nonsmokers, 647 men with lung cancer were cigarette smokers. In comparison, 27 case-control patients (i.e., patients in the hospital but with other ailments, not lung cancer) were nonsmokers, but 622 were cigarette smokers.
   A. What is the null hypothesis?
   B. Which is more appropriate: to calculate an odds ratio or to calculate an RRR?

C. This is a contingency table problem. Explain why
D. Conduct the test of the null hypothesis.
E. What is the value of the test statistic? Degrees of freedom? P-value?
F. Evaluate the results of your analysis — do you accept or reject the null hypothesis?

4. A more recent example from the study of the efficacy of St John's Wort as a treatment for major depression (Shelton et al 2001). Of patients who received St. John's Wort over 8 weeks, 14 were deemed improved while 98 did not improve. In contrast 5 patients who received the placebo were deemed improved while 102 did not improve.
   A. What is the null hypothesis?
   B. Which is more appropriate: to calculate an odds ratio or to calculate an RRR?
   C. This is a contingency table problem. Explain why
   D. Conduct the test of the null hypothesis.
   E. What is the value of the test statistic? Degrees of freedom? P-value?
   F. Evaluate the results of your analysis — do you accept or reject the null hypothesis?

9.3: Yates continuity correction

Introduction

Yates continuity correction: Most statistical textbooks at this point will note that critical values in their table (or any chi-square table for that matter) are approximate, but don't say why. We'll make the same directive — you may need to make a correction to the χ^2 for low sample numbers. It's not a secret, so here's why.

We need to address a quirk of the χ^2 test: the chi-square distribution is a continuous function (if you plotted it, all possible values between, say, 4 and 3 are possible). But the calculated χ^2 statistics we get are discrete. In our HIV-HG co-infection problem from the previous subchapter, we got what appears to be an exact answer for P, but it is actually an approximation.

We're really not evaluating our test statistic at the alpha levels we set out. This limitation of the goodness of fit statistic can be of some consequence — increases our chance to commit a Type I error — unless we make a slight correction for this discontinuity. The good news is that the χ^2 does just fine for most df values, but we do get concerned with its performance at $df = 1$ and for small samples.

Therefore, the advice is to use a correction if your calculated χ^2 is close to the critical value for rejection/acceptance of the null hypothesis and you have only one degree of freedom. Use the Yates continuity correction to standard χ^2 calculation, χ^2_c.

$$\chi^2_c = \sum_{i=1}^{k} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

For the 2×2 table (Table 9.3.1), we can rewrite the Yates correction:

$$\chi^2_c = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

Our concern is this: without the correction, Pearson's χ^2 test statistic will be **biased** (e.g., the test statistic will be larger than it "really" is), so we'll end up rejecting the null hypothesis when we shouldn't (that's a Type I error). This becomes an issue for us when the p-value is close to 5%, the nominal rejection level: what if p-value is 0.051? Or 0.049? How confident are we in concluding that we accept or reject the null hypothesis, respectively?

I gave you three examples of goodness of fit and one contingency table example. You should be able to tell me which of these analyses it would be appropriate to apply to correction.

More about continuity corrections

Yates suggested his correction to Pearson's χ^2 back in 1934. Unsurprisingly, new approaches have been suggested (e.g., discussion in Agresti 2001). For example, Nicola Serra (Serra 2018; Serra et al 2019) introduced

$$\chi^2_{Serra} = \frac{16}{N^3}(ad - bc)$$

Serra reported favorable performance when sample size was small and the expected value in any cell was 5 or less.

R code

When you submit a 2×2 table with one or more cells less than five, you could elect to use a **Fisher exact test**, briefly introduced here (see Section 9.5 for additional development), or, you may apply the Yates correction. Here's some code to make this work in R.

Let's say the problem looks like Table 9.3.1.

Table 9.3.1. Example 2×2 table with one cell with low frequency.

	Yes	No
A	8	12
B	3	22

At the R prompt type the following:

```
library(abind, pos=15)
#abind allows you to combine matrices into single arrays
.Table <- matrix(c(8,12,3,22), 2, 2, byrow=TRUE)
rownames(.Table) <- c('A', 'B')
colnames(.Table) <- c('Yes', 'No') # when you submit, R replies with the following ta
.Table  # Counts
 Yes No
A  8 12
B  3 22
```

Here's the χ^2 command; the default is no Yates correction (i.e., correct=FALSE); to apply the Yates correction, set `correct=TRUE`

```
.Test <- chisq.test(.Table, correct=TRUE)
```

Output from R follows:

```
.Test
    Pearson's Chi-squared test with Yates' continuity correction
data:  .Table
X-squared = 3.3224, df = 1, p-value = 0.06834
```

Compare without the Yates correction

```
.Test <- chisq.test(.Table, correct=FALSE)
.Test
    Pearson's Chi-squared test
data:  .Table
X-squared = 4.7166, df = 1, p-value = 0.02987
```

Note that we would reach different conclusions! If we ignored the potential bias of the un-corrected χ^2 we would be tempted to reject the null hypothesis, when in fact, the better answer is not to reject because the Yates-corrected p-value is greater than 5%.

Just to complete the work, what does the Fisher Exact test results look like (see Section 9.5)?

```
fisher.test(.Table)
    Fisher's Exact Test for Count Data
data:  .Table
p-value = 0.04086
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.9130455 32.8057866
sample estimates:
odds ratio
  4.708908
```

Which to use? The Fisher exact test is just that, an exact test of the hypothesis. All possible outcomes are evaluated and we interpret the results as likely as p=0.04086 if there is actually no association between the treatment (A vs B) and the outcome (Yes/No) (see Section 9.5).

Questions

1. With respect to interpreting results from a χ^2 test for small samples, why use the Yates continuity correction?
2. Try your hand at the following four contingency tables (a – d). Calculate the χ^2 test, with and without the Yates correction. Make note of the p-value from each and note any trends.

(a)

	Yes	No
A	18	6
B	3	8

(b)

	Yes	No
A	10	12
B	3	14

(c)

	Yes	No
A	5	12
B	12	18

(d)

	Yes	No
A	8	12
B	3	3

3. Chapter 9.1, Question 1 provided an example of a count from a small bag of M&Ms. Apply the Yates correction to obtain a better estimate of p-value for the problem. The data were four blue, two brown, one green, three orange, four red, and two yellow candies.

- Construct a table and compare p-values obtained with and without the Yates correction. Note any trend in p-value.

9.4: Heterogeneity chi-square tests

Introduction

After finding evidence to reject the statistical null hypothesis, it may be appropriate to proceed to test a number of data sets against the same theoretical distribution. Provided these are **planned comparisons**, part of the experiment and not an exercise in **data dredging** (Norman 2014), you may proceed to test a series of nested models where groups are combined to make new groups for comparison. This approach is analogous to the **post-hoc tests** one may conduct after a statistically significant ANOVA (see Chapter 12.4) or when identifying a **best-fit** regression model (see Chapter 18.4).

Example

Imagine a scenario where a population geneticist has collected allele (gene) frequency and genotype frequency data on single nucleotide polymorphisms (SNP) for the BRCA1 locus for three USA groups: African Americans, Asian Americans, and European Americans. Data of this kind can be obtained from the SNP database at NCBI (data retrieved 14 & 15 July 2014), and I collated several SNP involving a Cytosine base for Thymine base change (Table 9.4.1). The BRCA1 locus is on chromosome 17 and some of the hundreds of mutations found for this gene are associated with high risk of breast and or ovarian cancer (Couch and Weber 1996, Tram et al 2013).

Table 9.4.1. SNP of BRCA1 locus.

SNP	Population	n	C/C	C/T	T/T	C	T
rs1060915	African American	46	0.043	0.174	0,783	0.130	0.870
	Asian American	48	0.083	0.625	0.375	0.396	0.604
	European American	48	0.167	0.458	0.375	0.396	0.604
rs3737559	African American	40	0.043	0.174	0.783	0.130	0.870
	Asian American	48	0.083	0.625	0.292	0.396	0.604
	European American	48	0.167	0.458	0.375	0.396	0.604
rs799917	African American	124	0.048	0.258	0.694	0.177	0.823
	Asian American	90	0.467	0.444	0.089	0.689	0.311
	European American	118	0.407	0.458	0.136	0.636	0.364

We may wish to test the combined data set to see if the large data set differs from Hardy Weinberg expectations before proceeding with a series of sample populations. By combining the data sets, we will be able to test whether one genotype is different from exception.

The goal then is to pool the data so that you have a more powerful test of the null hypothesis (remember our general discussion of statistical power and how increasing sample size increases your chances of correctly rejecting the null hypothesis).

For example, in our test of Hardy-Weinberg expectations (Example C), we calculated a χ^2 test of 7.8955. This test has $k-1=3-1=2$ degrees of freedom. At alpha 5%, the critical value was 5.991. We clearly would reject the null hypothesis. But do we reject because of a difference in the **aa**, **ab**, or **bb** genotypes? Simply combine categories. Let's test the homozygotes (**aa** and **bb**) versus the heterozygotes (**ab**).

Table 9.4.2. Worksheet.

	Categories		
	aa + bb	**ab**	**n**
Observed	66	34	100
Expected	52	48	100
$\chi^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$	$=\frac{(66-52)^2}{52}+\frac{(34-48)^2}{48}=3.769+4.083=7.852$		

With one degree of freedom, we clearly reject the null hypothesis because the critical value at 5% and one degree of freedom is equal to 3.841. To get the p-value, use

```
pchisq(7.852,df=1,lower.tail=FALSE)
```

which will return

```
[1] 0.005076452
```

Oh, and what was the null hypothesis? That there was an equal number of homozygotes and heterozygotes.

Other tests would be possible here, but the point is that you can dissect an experiment to determine which group is causing you to reject the null original hypothesis. While this is an important tool, you should also consider issues of *a priori* and *a posterori* decisions in experiments.

Questions

1. Compare and contrast the purpose of Yates correction and heterogeneity chi-square test.

2. The SNP rs1799971 (A > G) in the μ-opioid receptor (MOR) is associated with opioid dependency. Allele frequencies for different populations are provided in the table. Global frequency for A at this SNP is 0.188 (therefore frequency of G is 0.812).

Population	n	A	G
African American	32510	0.97	0.03
Central American	2450	0.81	0.19
European	18872	0.86	0.14
Native American	1260	0.86	0.14
Native Hawaiian	4534	0.75	0.25
South Asian	856	0.59	0.41

- Combine the data sets and test whether genotype frequency is different from the reported global exception.
- Test each population separately.

9.5: Fisher exact test

Introduction

We mentioned that chi-square tests for contingency tables are fine as long as two conditions are met. These are the assumptions of a χ^2 test:

1. No cell should have expected values less than 5%.
2. The test performs poorly at DF = 1 because we are approximating an infinite distribution with an exact test.

You will note that any time you have a 2×2 table, the second condition is always an issue because 2×2 tables have $DF = 1$. Thus, in biomedical research, it is common to have an experiment that may be appropriate for a contingency analysis but the data may suffer from one or both of these limitations. **Fisher's exact test** is always an option for these types of problems, but with the advantage that it always returns the exact p-value.

As a reminder, the 2×2 table looks like:

Table 9.5.1. 2×2 table reporting numbers of subjects who have (Yes) or do not have (No) the event.

		Column 1	Column 2
	Subjects	Yes	No
Row 1	Treatment 1	**a**	**b**
Row 2	Treatment 2	**c**	**d**

where **a** is the count of Treatment 1-treated subjects who have the event, **b** is the count of Treatment 1-treated subjects who do not have the event, **c** is the count of Treatment 2-treated subjects who have the event, and **d** is the count of Treatment 2-treated subjects who do not have the event. Note the row and column totals:

$$\begin{aligned} \text{Row 1} &= a+b \\ \text{Row 2} &= c+d \\ \text{Column 1} &= a+c \\ \text{Column 2} &= b+d \end{aligned}$$

For example, a fairly common "Gee, that's curious" fact is that the seven left-handed US Presidents since 1901 (out of a total of 21 presidents) exceed the proportion of left-handers in the general population (about 10%). For comparison, we could ask the same question about Vice Presidents.†

Table 9.5.2. Left-handedness of US presidents and vice presidents since 1901.

Subjects	Yes	No
Presidents	7	14
Vice presidents	5	20

†Seven Vice-Presidents went on to become President, four right-handers, 3 left-handers.

Ronald A. Fisher came up with a test that is now called "Fisher's Exact test" that circumvents this problem. It is an extremely useful test to know about because it provides a way to get an exact probability of the outcome compared to all other possible outcomes. Thus, when asked for a possible alternate to the chi-square contingency test for a 2×2 table, you can respond "Fisher's Exact test."

Although tedious to calculate by hand and resource demanding when done by computer because of the multiple factorial expressions, the major advantage of the test is that it does not rely on the assumption that an underlying distribution applies. The Fisher Exact test can be used to calculate the exact probability of the observed outcome (P).

The equation for the Fisher Exact test can be written as

$$P = \frac{R_1! \cdot R_2! \cdot C_1! \cdot C_2!}{a! \cdot b! \cdot c! \cdot d! \cdot n!}$$

where R stands for row total, C stands for column total, n is the sample size, ! is the **factorial**, and a, b, c, and d are defined as in Table 9.5.1.

How does Fisher's Exact test work? The data are set up in the usual way for a contingency problem, but now, we calculate the probability for all possible outcomes that we COULD have seen from our experiment, and ask if the actual outcome is unusual (low p-value). The trick is recognizing that you have to keep the totals constrained (note row and column totals stay the same).

Table 9.5.3. Original 2×2 contingency table (bold), with the next two more extreme outcomes

original data	————>	more extreme	————>	next more extreme still	————>
Yes	**No**	Yes	No	Yes	No
10	**5**	11	4	12	3
4	**12**	3	13	2	14
	p-value=0.0206		p-value=0.0029		p-value=0.0002

I've shown just the one-tailed outcomes, so the p-values are for one-tailed tests of hypothesis. The essence of the test is to find all outcomes MORE extreme than the original, in one direction. The one-tailed P-value then is the sum of all probabilities from those more extreme tables of outcomes.

To get the two-tailed probability, remember that you multiply the one-tailed probability by two. More accurate methods are also available (Agresti 1992).

Calculation of Fisher's test involves using all possible combinations and factorials. Rcmdr has Fisher's 2×2 built in via the Contingency table and as part of some `Rcmdr` plugins (e.g., `RcmdrPlugin.EBM`, the Evidence Based Medicine plugin). Here we illustrate Fisher Exact test from the context menu in the main Statistics menu.

Alternatively, there are many web sites out there that provide an online calculator for Fisher's Exact test. Here's a link to one such calculator on GraphPad's web site, cookies must be enabled to run this calculator).

To get the Fisher Exact test, your data must already be summarized into a 2×2 table, in which case you can use

Rcmdr: Statistics → Contingency tables… → Enter and analyze two way table (then select Fisher's Exact test option).

	Smoker: No	Smoker: Yes
Vitamin use: No	14	26
Vitamin use: Yes	19	15

If the original data are available, do not tally the counts, let R do the work for you. The worksheet would be stacked like so. The image of the R worksheet below contains 4 columns: Sex (M/F), Smoker (Never, Former, Current), Smoke (Y/N), and Vitamin User (No, Regular).

Stacked worksheet for Continency table or Fisher exact test.

R code

To carry out contingency table analysis or Fisher Exact test,

Rcmdr: Statistics → Contingency tables… → Two way table …

Check the box next to the Fisher's exact test.

Select `Vitamin.Use` for Row variable and `Smoke` for Column variable. Click OK, and here is the R output.

```
> fisher.test(.Table)
Fisher's Exact Test for Count Data
data: .Table
p-value = 0.1008
```

```
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.1496417 1.1985775
sample estimates:
odds ratio
0.4302094
```

We accepted the defaults. Is this a one- or two-tailed test of hypothesis?

What can we conclude about the null hypothesis? Do we accept or reject?

Want to know what the “odds ratio” is? Follow the link to the next subchapter.

When to use the Fisher Exact Test?

Here’s the take-home message: the Fisher exact test is an alternate and better choice over the contingency table chi-square for 2×2 tables if one or more of the cells has expected values less than 5%. It is also appropriate for cases in which you have only 1 degree of freedom (as do all 2×2 tables!), but it doesn’t make sense if each cell has more than 5% expected values (the calculation is too tedious), but rather, apply the Yate’s correction. As the sample sizes get larger, the different methods converge to virtually identical answers.

Some examples.

Is there an association between final grades and attendance on a randomly selected day?

Table 9.5.4. First scenario

cc	Yes	No
Letter grade A	2	3
Other letter grade	1	6

Table 9.5.5. Second scenario.

cc	Yes	No
Letter grade A	5	6
Other letter grade	2	12

Table 9.5.6. Third scenario

cc	Yes	No
Letter grade A	10	12
Other letter grade	4	24

Code for tests are as follows for Table 9.5.4 as an example:

Data table:

```
grades.Table <- matrix(c(2,3,1,6), 2, 2, byrow=TRUE)
```

Chi-square test of independence:

```
.Test <- chisq.test(grades.Table, correct=TRUE)
```

Fisher Exact test:

```
fisher.test(grades.Table, alternative = “greater”)
```

Questions

1. Apply the Fisher exact test on the four contingency tables (a – d) introduced in Section 9.3, question 2. Make note of the p-value from Fisher exact test and from analyses used to complete question 2 in Section 9.3. Note any trends. (Hint: make sure you are testing the same null hypothesis.)

(a)

	Yes	No
A	18	6
B	3	8

(b)

	Yes	No
A	10	12
B	3	14

(c)

	Yes	No
A	5	12
B	12	18

(d)

	Yes	No
A	8	12
B	3	3

9.6: McNemar's test

Introduction

There are a number of scenarios in which subjects are **paired** or matched as part of the experimental design in order to control for confounding variables — a **matched pair case-control**. Subjects may be matched by age, or other criteria, or the observations are repeat measures of the same subjects (e.g., left hand vs. right hand). One member of each pair is then randomly assigned to a treatment, the remaining pair member then assigned to the other treatment group. This scenario should remind you of our standard contingency table problem, but instead of a random collection of subjects assigned to treatments, the data are paired nominal. Thus, paired means that experimental (sampling) units are not independent, which if ignored violates an assumption required to employ the χ^2 test. We use **McNemar's test** instead.

The possible results of such a design include just two outcomes: the pairs have the same outcome (agree, **concordant**) or the pairs have different outcomes (disagree, **disconcordant**).

McNemar's solution was to consider only the discordant pairs. Consider two kinds of tests or assays for a condition, and the doctor receives the results of both tests.

Table 9.6.1. Format of data where McNemar's test can be applied.

		Test 2		
		Positive	Negative	Row total
Test 1	Positive	a	b	$a+b$
	Negative	c	d	$c+d$
	Column total	$a+c$	$b+d$	n

Null hypothesis is that marginal proportions are equal:

$H_O = p_b = p_c$
$H_A = p_b \neq p_c$

Then McNemar's test is given by

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

and the test has one degree of freedom.

If one of the cells is low, then a continuity correction would be applied (Edwards 1948, cited in Fagerland et al 2013). With this correction the equation becomes

$$\chi^2_c = \frac{(|b-c|-1)^2}{b+c}$$

If either b or c is small, then the McNemar's test statistic does not approximate a χ^2 distribution very well, so there is a binomial version that you would use (Cochran's Q test) in cases where there are three or more matched sets and is common in meta-analysis (Kulinskaya and Dollinger 2015).

R code

Example data: Approval ratings for President Trump at two important markers during the Covid-19 pandemic: in April 2020, deaths passed 10,000 persons in the U.S.; in October 2020, it was reported that President Trump tested positive for SAR-COV2 and was admitted to Walter Reed National Military Medical Center (admitted 3 Oct., released 5 Oct.). Surveys were conducted by YouGov (April, sponsored by The Economist; October, sponsored by Yahoo News; data extracted from How Americans View Biden's Response To The Coronavirus Crisis)

Table 9.6.2. U.S. approval ratings for President Trump in 2020.

	Approve	Disapprove

	Approve	Disapprove
April survey	720	705
October survey	645	812

Enter the data as a matrix (note this would be a general approach for the contingency table problems, too, instead of entering via Rcmdr menu). The discordant pairs are $b = 645$ and $c = 705$.

```
covid19 <- matrix(c(720, 645, 705, 812), nrow = 2, dimnames = list("April survey" = c

 covid19
                     October survey
April survey       Approve   Disapprove
         Approve       720          705
      Disapprove       645          812
```

Uncorrected:

```
mcnemar.test(covid19, correct=FALSE)

McNemar's Chi-squared test

data: covid19
McNemar's chi-squared = 2.6667, df = 1, p-value = 0.1025
```

Correction applied:

```
mcnemar.test(covid19, correct=TRUE)

McNemar's Chi-squared test with continuity correction

data: covid19
McNemar's chi-squared = 2.5785, df = 1, p-value = 0.1083
```

Conclusions?

No change in approval ratings. The correction for small sample size had little effect on p-value, unsurprisingly, given that the surveys included 1500 (April) and 1504 (October) persons.

Unconditional paired tests

McNemar's solution considers only the discordant pairs; it's a conditional test. The downside of these tests is that the concordant pairs are not considered. Thus, by in effect tossing out some portion of the experimental results, it shouldn't surprise you that the statistical power of the test is reduced (see Chapter 11). Thus, McNemar's test may no longer be the best choice. Alternative unconditional tests have been proposed, and the **mid-P** alternative shows promise (Routledge 1994; Fagerland et al 2013). The mid-P value is calculated as the standard p-value for a test statistic minus one half the difference between the standard p-value and the next lowest possible p-value. McNemar's mid-p test is available in package `contingencytables` . Try with the example data set in Fagerland et al 2013 (Table 1).

```
#create a 2x2 matrix
bentur <- rbind(c(1, 1), c(7, 12))
```

First run McNemar's test without correction for small sample size.

```
mcnemar.test(bentur, correct=FALSE)
```

R output follows:

```
McNemar's Chi-squared test

data: bentur
McNemar's chi-squared = 4.5, df = 1, p-value = 0.03389
```

Next, run McNemar's test with correction for small sample size.

```
mcnemar.test(bentur, correct=TRUE)
```

R output follows:

```
McNemar's Chi-squared test with continuity correction

data: bentur
McNemar's chi-squared = 3.125, df = 1, p-value = 0.0771
```

Last, run mid-p version of McNemar's test.

```
McNemar_midP_test_paired_2x2(bentur)
```

R output

```
[1] The McNemar mid-P test: P = 0.039063
```

See also `mcnemarExactDP` function in `exact2x2` package. Without explanation, here's the R code and results.

```
mcnemarExactDP(n = sum(bentur), m= bentur[1,2] + bentur[2,1], x = bentur[1,2])

      Exact McNemar Test (with central confidence intervals)

data: n=sum(bentur) m=bentur[1, 2] + bentur[2, 1] x=bentur[1, 2]
n = 21, m = 8, x = 1, p-value = 0.07031
alternative hypothesis: true difference in proportions is not equal to 0
95 percent confidence interval:
 -0.54549962 0.02044939
sample estimates:
       x/n    (m-x)/n  difference
0.04761905 0.33333333 -0.28571429
```

Alternatively, use wrapper function `mnemar.exact()` .

```
mcnemar.exact(bentur)
```

R output:

```
Exact McNemar test (with central confidence intervals)

data: bentur
b = 1, c = 7, p-value = 0.07031
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.003169739 1.111975554
sample estimates:
odds ratio
 0.1428571
```

Note the alternative hypothesis: p-value is two-tailed.

Questions

1. Apply McNemar's test and mid-P exact test to CDC example

		Controls	
		Exposed	Not exposed
Cases	Exposed	58	89
	Not exposed	32	95

9.7: Chapter 9 References and Suggested Readings

Agresti A (1992) A Survey of Exact Inference for Contingency Tables. *Statistical Science* 7(1):131-153.

Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in medicine* 20(17-18): 2709-2722.

Apaydin, E. A., Maher, A. R., Shanman, R., Booth, M. S., Miles, J. N., Sorbero, M. E., & Hempel, S. (2016). A systematic review of St. John's wort for major depressive disorder. *Systematic reviews*, 5(1), 148.

Bellinger, D., Leviton, A., Waternaux, C., Needleman, H., Rabinowitz, M. (1987). Longitudinal Analyses of Prenatal and Postnatal Lead Exposure and Early Cognitive Development. *New England Journal of Medicine* 316:1037-1043.

Bewick, V., Cheek, L., Ball, J. (2003). Statistics review 8: Qualitative data — tests of association. *Critical Care* 8:46-53.

Bhattarai, N., Stapleton, J. T. (2012). GB virus C: the good boy virus? Trends in microbiology 20:124-130.

Cohen, J. (1992). Statistical power analysis. *Current directions in Psychological Science* 1:98-101.

Cook, N. R., Lee, I.-M., Zhang, S. M., Moorthy, M. V., Buring JE (2013) Alternate-Day, Low-Dose Aspirin and Cancer Risk: Long-Term Observational Follow-up of a Randomized Trial. *Annals of Internal Medicine* 59:77-85.

Couch, F. J., Weber, B. L. (1996). Mutations and Polymorphisms in the familial early-onset breast cancer (BRCA1) gene. *Human Mutation* 8:8-18.

Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, *13*(1), 91.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical; Society. Series B (Methodological)* 17:69-78.

Graffelman, J., & Weir, B. S. (2018). On the testing of Hardy-Weinberg proportions and equality of allele frequencies in males and females at biallelic genetic markers. *Genetic Epidemiology*, 42(1), 34-48.

Guidetti, P. (2006). Marine reserves reestablish lost predatory interactions and cause community changes in rocky reefs. *Ecological Applications* 16:963-976.

Hrastnik, D., Budija, F., Humar, M., & Petrič, M. (2013). Influence of liquefied and CCB containing liquefied wood on growth of wood decay fungi. *Maderas. Ciencia y tecnología*, *15*(1), 105-118.

Kroonenberg, P. M., and Verbeek, A. (2018). The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do? *The American Statistician*, 72(2), 175-183.

Kulinskaya, E., & Dollinger, M. B. (2015). An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. *BMC medical research methodology*, 15, 1-19.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American statistical Association*, *88*(424), 1242-1249.

Norman, G. (2014). Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. *Advances in Health Sciences Education* 19(1):1-5.

Routledge, R. D. (1994) Practicing Safe Statistics with the Mid-p. *The Canadian Journal of Statistics* 22(1):103-110.

Serra, N. (2018). A significant minimization of Pearson's X2 statistics in 2×2 contingency tables: preliminary results for small samples. *Epidemiology, Biostatistics, and Public Health*, 15(3).

Serra, N., Rea, T., Carlo, P. D., & Sergi, C. (2019). Continuity correction of Pearson's chi-square test in 2×2 Contingency Tables: A mini-review on recent development. *Epidemiology, Biostatistics, and Public Health 16*(2), Article 2. https://doi.org/10.2427/13059

Shelton et al. (2001). Effectiveness of St John's Wort in Major Depression: A randomized control trial. *Journal of the American Medical Association* 285:1978.

Tram, E., Savas, S., Ozcelik, H. (2013). Missense variants of uncertain significance (VUS) altering the phosphorylation patterns of BRCA1 and BRCA2. *PLoS One* 8(5):e62468.

Vickers, A. J. (2009). What is a p-value anyway? 34 stories to help you actually understand statistics. Pearson College Division.

Whitley, E., Ball, J. (2002). Statistics review 3: Hypothesis testing and P values. *Critical Care* 6:222-225

Xiang, J., Wünschmann, S., Diekema, D. J., Klinzman, D., Patrick, K. D., George, S. L., Stapleton, J. T. (2001). Effect of Coinfection with GB Virus C on Survival among Patients with HIV Infection. *New England Journal of Medicine* 345:707-714.

Xu, B., Feng, X., Burdine, R. D. (2010). Categorical data analysis in experimental biology. *Developmental Biology* 348(1):3-11.

Yates (1984). Tests of Significance for 2 × 2 Contingency Tables. *Journal of the Royal Statistical Society Series A: Statistics in Society* 147(3):426-449.

CHAPTER OVERVIEW

10: Quantitative Two-Sample Tests

Introduction

A one-sample parametric test compares the mean against a population value. The population value may come literally from census information, or, more likely, it comes from some applicable theory. The one-sample t-test was presented, along with how to calculate the confidence interval, in the previous chapter.

In this chapter, we also extend to considering two-sample tests, about hypotheses for two groups. The two groups may consist of observations on different subjects, and thus the two groups are independent of each other — an **independent sample t-test** may be used to test null hypothesis. A common experimental design is to measure individuals two or more times, e.g., observations like body mass index, BMI, recorded on individuals at the start of an exercise program, and again on the same individuals some time after a treatment — a repeated measures design. In this case, the measures are paired and are, thus, not independent, and a **paired-sample t-test** would be advised.

Two-sample parametric tests are used to answer questions about the mean where the data are collected from two random samples of independent observations, each from an underlying normal distribution. The samples may be independent or paired, in which different hypotheses are tested.

10.1: Compare two independent sample means

Introduction

We introduced the concept of comparing a sample statistic (mean) against a population parameter (Chapter 6.7, **Normal deviate**) or **one-sample t-test** against a specified mean (e.g., from published data or from theory, Chapter 8.5).

Consider now a basic experimental design, the **randomized control trial**, or **RCT** (Fig. 10.1.1), introduced in Chapter 2.4.

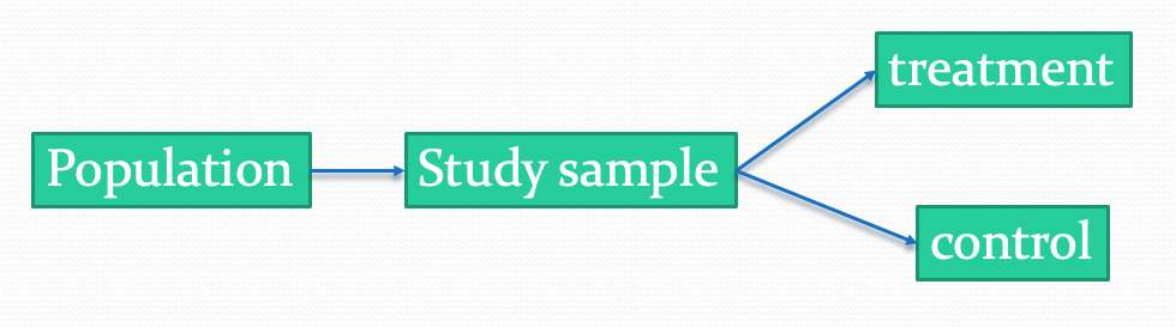


Figure 10.1.1: A two-group Randomized Control Trial.

Subjects **randomly selected** from population of interest, then again — **random assignment** — once recruited into one of two treatment groups. Importantly, subjects belong to one treatment arm only: no subject simultaneously receives the treatment and the control. This is in contrast to the paired design, in which subjects receive both treatments (see Chapter 10.3).

In inferential statistics about an experiment, we are more likely trying to test if sample means are different. For example:

- two species grown in a common garden, do they differ in growth rate?
- human subjects given a new treatment have better outcomes compare to those receiving a control treatment (e.g., placebo).

The equivalent null hypothesis is that two samples are pulled from the same population. We write the null hypothesis as $H_O : \bar{X}_1 = \bar{X}_2$

and the corresponding alternate hypothesis, H_A, then must be $H_A : \bar{X}_1 \neq \bar{X}_2$.

Question: Is this a one-tailed or two-tailed hypothesis?

Answer Two-tailed (review Chapter 8.4)

> **Note:**
>
> Note that in this day and age, there's really no reason to learn the t-test. First, it is just a special case of the one-way ANOVA; therefore, it's a special case of the general linear model. Struggling to learn R commands? Well, one solution would be just to learn the general linear model approach — just learn the R function `lm()` (OK, don't get too excited — `lm()` has many options and details). Second, few experiments or observational studies are likely to have only two groups; thus, the temptation to carry out a series of t-tests, taking all groups two at a time, or "pairwise," while tempting, actually violates a whole bunch of basic statistical rules (discussed in Chapter 12.1). It will also make statisticians go crazy when they see it. That said, if your experiment has but two groups, then by all means, the t-test is a choice. The t-test is also a statistical test that you have likely already used so we present the discussion here to build on what you may already have learned. We also present the independent t-test as a vehicle.

Worked example

We introduce the two-sample t-test, or better, the independent sample t-test.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

where the numerator is the difference between the two sample means and the denominator is the standard error of the differences between the two groups' standard errors. The formula for this standard error is

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The choice of independent sample over two-sample is best because it emphasizes that the two groups (the two samples), must be comprise of **independent sampling units**. This is a pretty straight-forward requirement; you have randomly assigned twenty individuals to two groups, a control group $(n = 10)$ and a treatment group $(n = 10)$. Individuals are either in the control group or they are in the treatment group — they cannot simultaneously appear in both groups.

We will work our way through this test by example. For starters, let's use the same lizard dataset (see **Example data set**, below), four body mass recordings (grams) each for house geckos (*Hemidactylus frenatus*, Fig. 10.1.2) and the Carolina anole (*Anolis carolensis*, Fig. 10.1.3), two of many lizard species introduced to Hawaii.

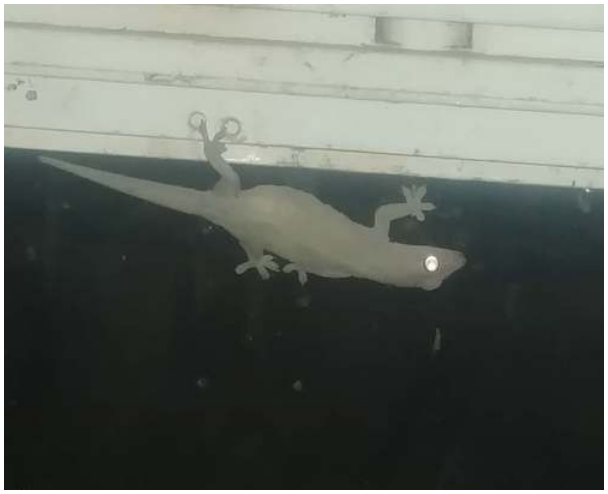


Figure 10.1.2: Male *Hemidactylus frenatus*, central Oahu.

Figure 10.1.3: Male *Anolis carolinensis*, `Akaka Falls, Big Island of Hawaii.

Example data set

```
Geckos: 3.186, 2.427, 4.031, 1.995

Anoles: 5.515, 5.659, 6.739, 3.184
```

Question: How would you go about creating a data frame with the values in long form (**stacked worksheet**), including a label variable and the body mass?

> **Note:**
>
> This test in `Rcmdr` requires that data are in **stacked worksheet** form in two columns, and not in **unstacked worksheet**. If you need help with worksheet format, then see Part07 in Mike's Workbook for Biostatistics.

Answer At the R prompt, type

```
Geckos <- c(3.186, 2.427, 4.031, 1.995); Anoles = c(5.515, 5.659, 6.739, 3.184) #crea
bmass <- c(Geckos, Anoles)    #combine the two vectors into a single vector holding a
species <- c("gecko", "gecko", "gecko", "gecko", "anole", "anole", "anole", "anole")
lizards <- data.frame(species, bmass)  #create your data frame
lizards   #print your data frame  species bmass
1   gecko 3.186
2   gecko 2.427
3   gecko 4.031
4   gecko 1.995
5   anole 5.515
6   anole 5.659
7   anole 6.739
8   anole 3.184
```

Note also that you can enter data into the Data editor by creating the data frame first then adding values. To edit the data frame "lizards" type `fix(lizards)` at the R prompt, then close the data frame when you have added or changed values as needed.

As always, begin with an exploration of the data, including a graph (Fig. 10.1.4).

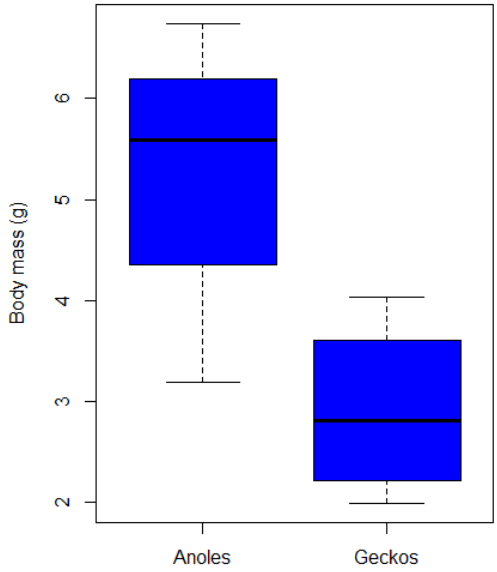


Figure 10.1.4: Box plot of lizard body mass.

We can see already that there's greater spread of data for the Anoles compared to the Geckos, but the median values differ. Small sample sizes can be a problem for analyses as we can only have reduced confidence in our conclusions. However, we press on for the sake of demonstration.

Let's test the null hypothesis, H_O, i.e., the two species of lizards have the same mean body mass.

Rcmdr: Statistics → Means → Independent-samples t-test…

In this next image I posted the Rcmdr menu popup for the Independent Samples t-test. Later versions of Rcmdr split the settings for this command into two tabs; the first tab allows for the selection of the variables and setting the hypotheses whereas the second tab, labeled Options, permits additional choices. The default selections need your attention: to actually conduct the t-test you need to answer "No" to the question, "Assume equal variance?"

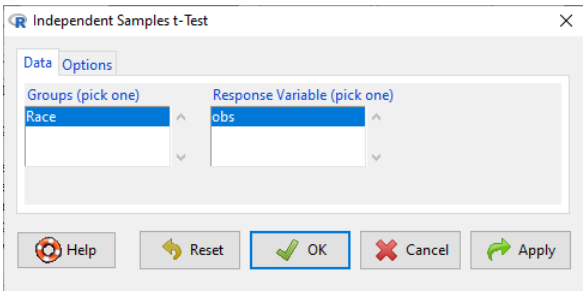


Figure 10.1.5: Rcmdr Data menu for Independent sample t-test.

Select the Options tab (Fig. 10.1.6) to select null hypothesis and to select the t-test and not the **Welch-test** (which is the default, i.e., No to the prompt "Assume equal variances?").

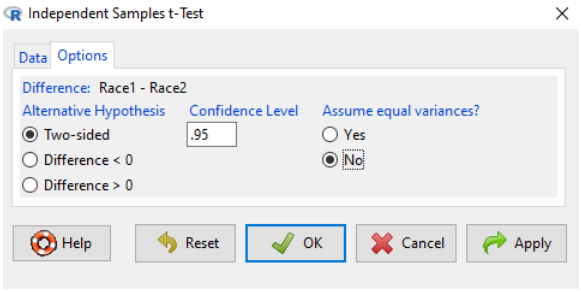


Figure 10.1.6: Rcmdr Options menu for Independent sample t-test.

Let's look at the results and break down the parts of the test.

```
t.test(Body.mass~Lizard, alternative='two.sided', conf.level=.95,
+ var.equal=TRUE, data=lizards)

Two Sample t-test

data: Body.mass by Lizard
t = 2.7117, df = 6, p-value = 0.03503
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.2308685 4.4981315
sample estimates:
mean in group Anolis mean in group Gecko
5.27425 2.90975
```

Consider the R session output above and answer the following questions.

Questions for the worked example

1. Which lizard group had the greater mean value, *Anolis* or *Gecko*?
2. What are the assumptions necessary for you to use the independent sample t-test?
3. What does "two-sided" mean?
4. What was the null hypothesis?
5. Was this a one-tailed or two-tailed test of the null hypothesis?
6. What is the value of the test statistic?
7. How many degrees of freedom?
8. What is the critical value for this test?
9. What is the value of the lower limit of the 95% confidence interval?
10. What is the value of the lower limit of the 99% confidence interval?

11. True or False. If the null hypothesis is accepted, then zero is a value included in the 95% confidence interval.
12. Do you accept the null hypothesis? Explain your selection.

Try another example

DNA damage, changes in the chemical structure of nucleotide bases or breakage of the DNA chains, occurs in cells under many circumstances. The comet assay, or single-cell gel electrophoresis, is one method for visualizing and measuring DNA strand breaks in cells. Exposed cells are mixed with a low-melting temperature agarose and placed onto a microscope slide. The cells are then lysed with an alkaline detergent and high salts. When current is applied across the slide, undamaged DNA remains in the nucleus, whereas damaged DNA extends towards the anode to form a comet-like tail, with imaging assisted by including a fluorescent dye like Sybr-Green. Examples of comets are shown below (Fig. 10.1.7).

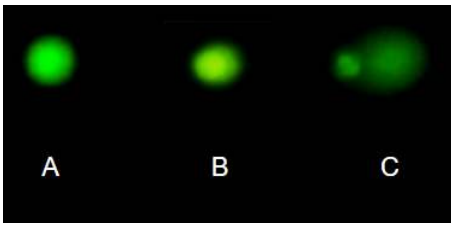


Figure 10.1.7: Comet examples. A: Intact cell, no DNA damage, B: Cell with some DNA damage, a slight tail to the right is evident, C: Cell with significant DNA damage, a large tail is evident.

In an experiment, immortalized lung epithelial cells were exposed to dilute copper solutions for 30 minutes, then washed with PBS. The comet assay was applied to these cells and for comparison, to cells without copper exposure but otherwise treated the same way (controls). The data are available at the bottom of this page (scroll down or click here).

Again, you should begin all analyses with an exploration of the data, including a graph (Fig. 10.1.8).

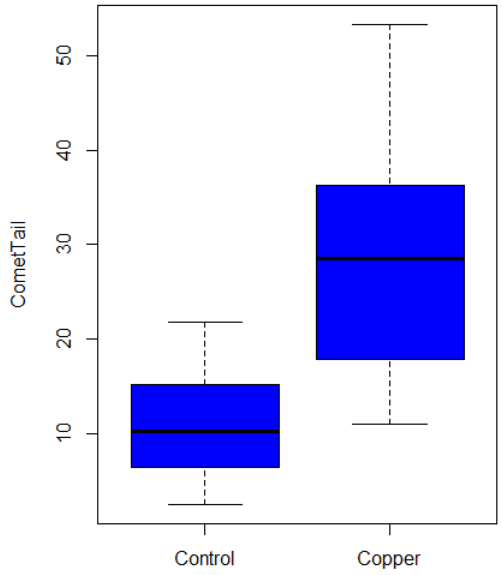


Figure 10.1.8: Boxplot of comet tail lengths for cells with and without (control) exposure to copper in the cell medium for 30 minutes.

Let's look at the R output for the t-test analysis.

```
    Two Sample t-test

data:  CometTail by Treatment
t = -5.8502, df = 38, p-value = 9.139e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.39865 -10.88213
```

```
sample estimates:
mean in group Control  mean in group Copper
             11.14533              27.78571
```

Consider the R session output above and answer the following questions.

Questions for Comet assay data set

1. Which cell group had the greater mean value, Copper-exposed or Control-exposed cells?
2. What are the assumptions necessary for you to use the independent sample t-test?
3. What does “two-sided” mean?
4. What was the null hypothesis?
5. Was this a one-tailed or two-tailed test of the null hypothesis?
6. What is the value of the test statistic?
7. How many degrees of freedom?
8. What is the critical value for this test?
9. What is the value of the lower limit of the 95% confidence interval?
10. What is the value of the lower limit of the 99% confidence interval?
11. True or False. If the null hypothesis is accepted, then zero is a value included in the 95% confidence interval.
12. Do you accept the null hypothesis? Explain your selection.

T test from summary statistics

In some cases you may only have to summary statistics for data, e.g., the means and the standard deviations. We can use the equations of the t test to write a simple formula, where the user provides the known means, standard deviations, and sample size. For example, create a simple function with `readline` for user input.

```
myTtest <- function() {
   mnx <- as.numeric(readline(prompt="Enter mean of x: "))
   stdevx <- as.numeric(readline(prompt="Enter sd of x: "))
   nx <- as.numeric(readline(prompt="Enter n of x: "))
   mny <- as.numeric(readline(prompt="Enter mean of y: "))
   stdevy <- as.numeric(readline("Enter sd of y: "))
   ny <- as.numeric(readline(prompt="Enter n of y: "))
   myTvalue <- abs(((mnx-mny)-0)/sqrt(((stdevx^2)/nx)+(stdevy^2)/ny))
   myDF <- as.integer(nx+ny-2)
   myPvalue <- pt(myTvalue,myDF,lower.tail=FALSE)*2
   myResults <- c(myTvalue, myDF, myPvalue)
   report <- c("T-test: ", "df: ", "two-tailed p-value: ")
   cat(sprintf("%s %3.3f, ", report, myResults))
}
```

then run the function by typing `myTest()` at the R prompt and entering the means, standard deviations, and sample size when prompted.

```
myTtest()
Enter mean of x: 2.91
Enter sd of x: .895
Enter n of x: 4
Enter mean of y: 5.27
Enter sd of y: 1.497
```

```
Enter n of y: 4
T-test: 2.706, df: 6.000, two-tailed p-value: 0.035
```

Questions

1. Don't forget to work through the Questions for the Comet tail data set (scroll up or click here).
2. Microsoft Excel, LibreOffice Calc, and Google sheets spreadsheet software all include t-test functions and return the p-value. Consider two variables big (100, 110, 120, 100, 110, 210, 200) and small (0,1,1,2,0,1,0). (Note — these two groups are obviously very different, calculating a t-test on their difference is silly, just for this question.) If formatting is set to the default two decimal places for Number cell category, the p-value will return as "0.00." How should you report the p-value in this case?
3. For the t-test, and in general for reporting of all statistical tests, what three numbers reported in the R output should you minimally report?

Comet assay data set

Treatment	CometTail
Control	17.856139
Control	16.52125
Control	14.925449
Control	14.029174
Control	13.332945
Control	8.811185
Control	14.701654
Control	9.261025
Control	21.779311
Control	6.180284
Control	9.201752
Control	5.54472
Control	6.717885
Control	2.625092
Control	7.191583
Control	5.392866
Control	11.284813
Control	15.441254
Control	17.857176
Control	4.250956
Copper	53.214287
Copper	38.92857
Copper	18.928572
Copper	30
Copper	28.928572

Copper	15.357142
Copper	17.857143
Copper	17.5
Copper	21.071428
Copper	29.285715
Copper	28.214285
Copper	16.785715
Copper	21.071428
Copper	37.5
Copper	38.214287
Copper	17.857143
Copper	29.642857
Copper	11.071428
Copper	35
Copper	49.285713

10.2: Digging deeper into t-test plus the Welch test

Introduction

We need to spend some more time with the **independent sample t-test**; by tearing it apart, we can learn about how **parametric tests** work in general.

Our assumptions for the independent sample t-test are like the one-sample t-test the data must be continuous and **normally distributed** (one of our standard assumptions for parametric tests, see Chapter 13). The formula is very similar to the one-sample t-test, except that now we have two sample means and the formula for the **standard error** (SE) has also changed.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

We see that the test statistic t is large if the numerator is large compared to the denominator. Large values of t will be evidence in favor of rejecting the null hypothesis.

The numerator is straight-forward: we subtract one sample mean from the other — if there is no difference between the samples, then this difference will be close to zero.

The denominator requires additional discussion. What is it called? It is the **pooled standard error of the mean** (pooled SEM). Provided the assumption of equal variances holds — an additional standard assumption for parametric tests, see Chapter 13 — then sample variances estimate the population variance and we can use this information to our advantage to best test the hypothesis about the sample means. In other words, we don't have to lose a degree of freedom to account for differences in variability for the two groups (see Welch test). More degrees of freedom means more statistical power to test the null hypothesis and at the same time, more confidence that the test is performing to its best.

Let's break down the pooled standard error of the mean in order to see how the assumption of equal variances affects the t-test. We assume that $s_1^2 = s_2^2$.

Recall that $H_O : \bar{X}_1 = \bar{X}_2$.

Now we want a "pooled SE" that is the pooled standard error for both samples. The variance of the difference between the means can be written as

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

First we need to calculate the pooled variance, where v_1 and v_2 are degrees of freedom for sample one and sample two, respectively. Note that this is just simply a combined formula of the sample variance.

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2}$$

What is SS? It is the sum of squares, where SS_1 and SS_2 refer to sum of squares for each of the two groups.

$$SS = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

You should recognize SS from your definition of the sample variance, Chapter 3.2.

And the **standard error for the difference** between two means can now be written as

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

so the 2-sample t-test can be written to reflect the pooled sample standard error of the difference between two sample means. We can see how **unequal sample size** is accommodated by the t-test.

Conducting the test by hand follows the same form as the one-sample t-test. Find the degrees of freedom (DF), but now for each sample.

Finally we evaluate with the critical value in Table (Appendix, Table of Student's t distribution) and compare the **t test statistic** against the critical value with the appropriate degrees of freedom.

Because this is the t-test, again, we are assuming that the variances are the same between the two populations (**homoscedasticity**) and this allows us to pool the variance. As it turns out, the T-test is not overly sensitive to other deviations from the assumptions, but if the variances are in fact different, then the standard formula may yield incorrect Type I error rates compared to stated probability level (α).

However, it would be poor statistical choice to use a test where there are alternatives. This is why in part that R (`Rcmdr`) sets as the default for the t-test that variances are unequal! In fact, R does not do a t-test unless you change the default to assume equal variances, which, as we now know, is the t-test.

Welch test

What to do when assumptions for the t-test are not met? Many options have been proposed, and **Welch's approximate *t*** is a good alternative to the two-sample t-test — it would be appropriate if the normal assumption still held.

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom for the Welch's test are now

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot v_1} + \frac{s_2^4}{n_2^2 \cdot v_2}}$$

Note that all the Welch test does is remove the pooled estimate of the standard error, replaced with both variance estimates directly.

As a default option, R and Rcmdr uses a variation of Welch's test when you select to do the t-test without making the assumption of equal variance.

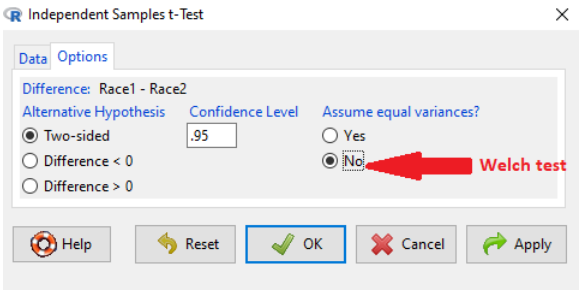


Figure 10.2.1: Screenshot Rcmdr t-test options. Default is "No" for Assume equal variances, i.e., the Welch test.

The Welch test is not a nonparametric test, it is a different formulation of the t-test.

Justification for beginning with t-test

It's unlikely that you will need the t-test in today's research climate. Data sets are large, experiments are complex with multiple variables and samples. Why do we have to consider the t-test, and then a separate test in the case for unequal sample size? I view it as a teaching moment. It makes the general point that ALL statistical tests make assumptions about how the calculations are done and as to the nature of the data set.

This is our first experience with what to do if there is a violation of an assumption of parametric tests (see Chapter 13). Here, the assumption is that the two groups have equal sample size. When they do not, the standard t-test tends to **biased estimates**. On the other hand, if the assumptions are met, the standard test is the best test because it has more power to do what we intended — that is, it is best at the actual test of the null hypothesis! Take heart — this point is not always appreciated even by scientists (cf. Fagerland 2012).

Let's us approach the problem of violation of assumptions in a couple of ways as an introduction to how, in general, to approach choice of statistical tests.

1. **Power of a test.** Much current statistical research focuses on learning about how a particular statistical test works when assumptions are violated. Thus, in addition to learning what tests are designed to do, we need to consider the effects of violations of assumptions on the performance of the test (namely, is the Type I error at the stated alpha level?). This is a matter of statistical power; power of a test reflects how well the test is able to get you the correct result even if assumptions are violated. Often tests perform well if sample sizes are large, despite violation of assumptions. We mention without proving that the two-sample t-test is robust to violations of normality assumption, to lack of equal sample sizes, and even to unequal variances. But good experimental design attempts to meet the assumptions because the test does better!
2. **Alternate forms of some tests are available to handle some aspects of test violations**. For example, the simple two-sample t-test can be modified to accommodate different variances (Welch's formula). Or you must find a different test (e.g. nonparametric tests).

In conclusion, all tests begin with consideration of the assumptions. In some cases we can test our assumptions. For example, we learned about testing the assumption of normal distributions of sample data. We can also test the assumption of equal variances.

Questions

1. Consider a clinical trial in which resting blood pressure is recorded on hypertensive subjects at the start of the trial, then 6 weeks after subjects have received daily supplements of flaxseed. Thus, for each subject there are two measures of blood pressure, BEFORE and AFTER.
   - Write the null hypothesis.
   - Write the alternative hypothesis.
   - Justify why or why not the hypothesis should be two-tailed. Explain why or why not an independent sample t-test may be used to compare.
2. Justify why the default t-test in R and therefore Rcmdr applies the Welch test, not the t-test?

10.3: Paired t-test

Introduction

Good experiments include controls. Interested in a new treatment for weight loss? Define a control group to compare the weight loss by a group using the new product. In many cases, the best control is the individual.

Consider now a basic experimental design, the **randomized crossover trial** (Fig. 10.3.1), introduced in Chapter 2.4.

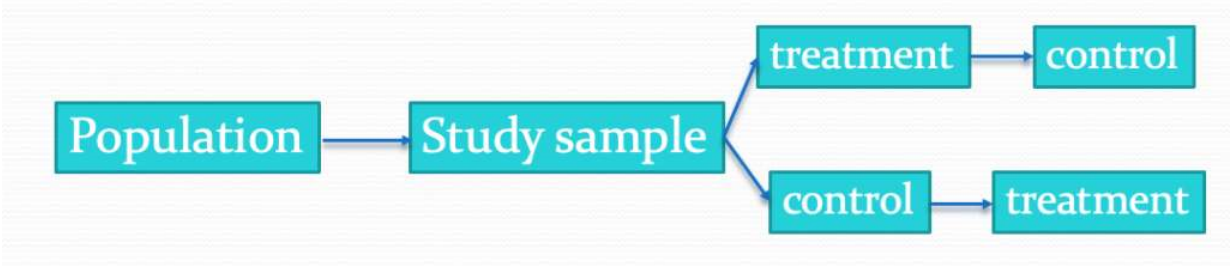


Figure 10.3.1: A two group Randomized Crossover Trial.

Subjects are randomly selected from a population of interest, then again once recruited into one of two **treatment arms**: arm 1, subjects first receive the experimental treatment, then some time later the subjects receive the control treatment; arm 2, subjects first receive the control treatment, then some time later the subjects receive the experimental treatment. Note the difference between this **paired** or **repeated measures design** and the independent sample design (see Chapter 10.1). Repeated measures designs have many advantages; we discuss them further in Chapter 14.6. At the start, repeated measures designs have greater statistical power compared to **cross-sectional (independent) sample designs.**

Many experiments are designed so that subjects receive all treatments and responses are gauged against the initial values recorded on the subjects. Repeated measures statistical tests, like the paired t-test, are needed however to analyze the data. These types of statistical procedures are similar to the two-sample independent t-test that we discussed earlier.

However, there is an important difference between these two types of statistical procedures. For the two-sample independent t-test the samples are unpaired: we observed one variable on some individuals assigned to two different groups. These groups might be

- Two locations where we measure plants or animals
- A treatment (or experimental) group with a control group.
- Expression of cytokeratin genes (e.g., $\Delta\Delta C_T$, fold-change) from breast cancer patients compared to healthy donor subjects (Andergassen et al 2016).

The point is that samples in one group are not the same samples in the second group.

In the paired t-test we have two groups, but the observations in these two groups are paired. Paired means that there is some relationship between one observation in the first sample and one observation in the second sample (every observation in one sample must be paired with one observation in another sample).

For example, weight in humans before and after a change in diet could be performed as a paired analysis. Each subject's weight before the diet was "paired" with the same subject's weight after the diet.

Another example comes from genetics. Siblings or monozygotic twins or clones, strains or varieties of plants or animals can be paired in an experiment.

- You can give one of the twins a particular diet, or the plant or animal clones or strains can be raised in a particular environment (nutrient)
- The other twin or plant or animal clone or variety can serve as a type of control by providing a normal diet or normal environment.

Another example is a study of environmental pollution on cancer rates in many different communities.

- The researchers selects pairs of communities with similar characteristics for many socioeconomic factors.
- Each pair of communities differed with respect to the proximity to a known source of pollution: one of the pair was close to a source of pollution and one of the pair was far from a source of pollution.

The purpose of pairing in this example is to attempt to "control" for all the socioeconomic factors that might contribute to cancer but they did not want to directly measure. These other factors should be similar for each member of the pair.

Example: How repeatable is human running performance?

Here's an example in which a measure was taken twice for the same individuals. The data are running speed or pace during a 5K race held annually on Oahu for a random sample of female runners (20 – 29 years old). The race was run annually on Oahu, and the data reported are the pace for the first race and the second race, which occurred a year later (Jamba Juice – Banana Man Chase, Ala Moana Beach Park, data extracted from source, https://timelinehawaii.com).

Table 10.3.1. 5K pace times (kph) for 15 women (20 – 29 years).

ID	Race 1	Race 2
1	15.28	15.61
2	11.22	11.19
3	8.80	9.14
4	8.88	5.46
5	9.81	10.50
6	6.12	5.69
7	8.31	8.71
8	6.26	7.42
9	17.16	16.41
10	16.23	15.82
11	5.90	7.12
12	8.31	10.48
13	5.93	8.64
14	10.54	5.99
15	9.53	8.69

Load the data into R as an unstacked data set. Data available at end of this page or click here.

Begin with description and exploration of the data. Start with **histograms** to get a sense of the sample distributions (hint: we're looking to see if the data looks like it could come from a normal distribution, see Chapter 13.3: Assumptions).

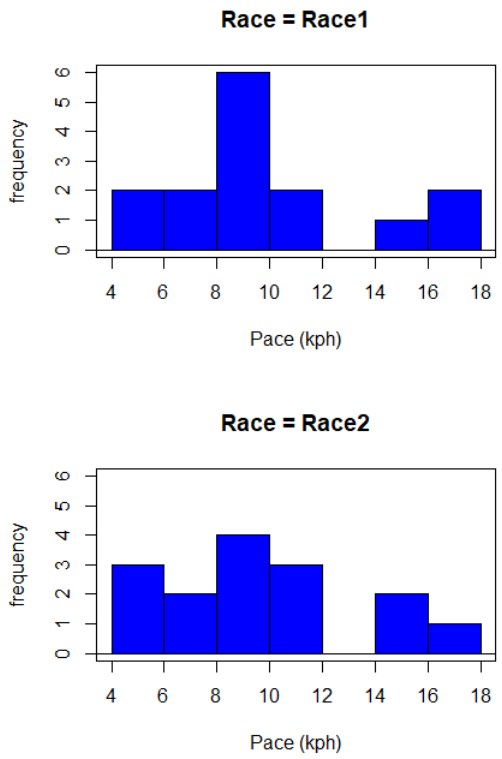


Figure 10.3.2: Histograms show the distribution of 5K running paces of 15 women who ran the race twice.

R code (stacked data set, then used defaults R Commander to make the histogram, then modified the code and submitted modified code to make Fig. 10.3.2)

```
with(stackExCh10.3, Hist(obs, groups=Race, scale="frequency", breaks="Sturges", col="
xlab="Time (min)", ylab="Frequency")))
```

Conclusion? The histograms don't look normally distributed so we keep this in mind as we proceed.

Here is a box plot comparing the first and second pace times.

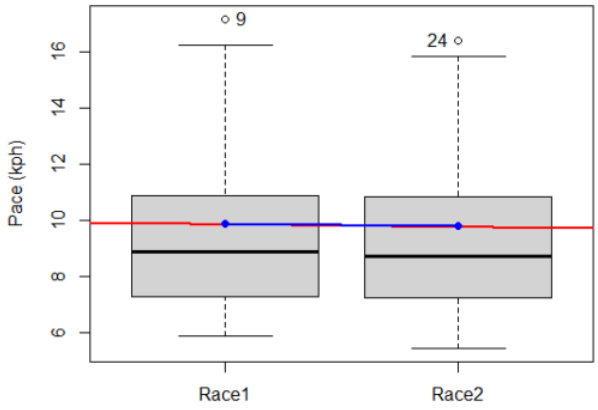


Figure 10.3.3: Box plot of race speed (kph) for 15 women 5K in two successive years.

I added a red trend line (linear regression, see Chapter 17) and connected the averages (blue line) for visual emphasis that there are no differences between the means, but note that one wouldn't do this as part of an analysis (see Chapter 4 discussion).

R code for Fig. 10.3.3

```
Boxplot(obs~Race, data=stackExCh10.3, id=list(method="y"), xlab="", ylab="Pace (kph)"
abline(lm(obs ~ as.numeric(Race), data=stackExCh10.3), col="red", lwd=2)
means <- tapply(obs, Race, mean)
points(1:2, means, pch=7, col="blue")
lines(1:2, means, col="blue", lwd=2)
```

The box plot works to show the median difference, but loses the paired information. A nice package called `PairedData` has several functions that work well with paired data.

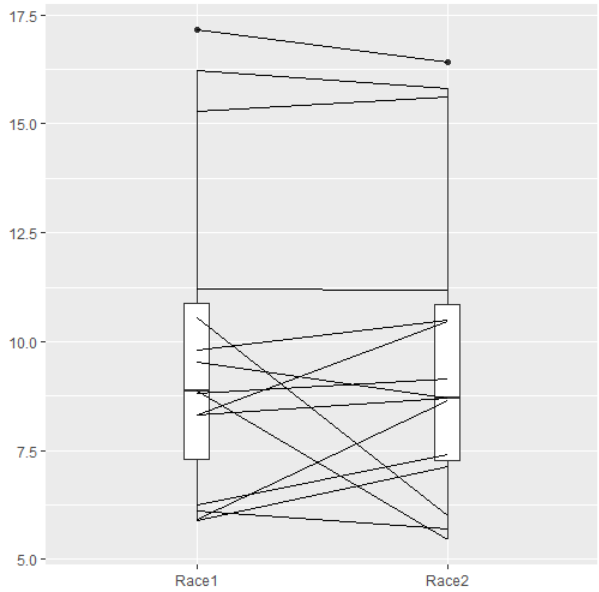


Figure 10.3.4: Profile plot, PairedData package.

R commands for Fig. 10.3.4:

```
require(PairedData)
attach(example.ch10.3) # remember to attach dataframe so you don't have to call varial
races <- paired(Race1, Race2)
plot(races, type = "profile")
```

Paired t-test calculation

The paired t-test is a straight-forward extension of the independent sample t-test; the key concept is that the two samples are no longer independent, they are paired. Thus, instead of mean of group 1 minus mean of group two, we test the differences between sample 1 and sample 2 for each paired observation.

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

1. Compute the differences between the Paired Samples (as in tables above)
2. Calculate the MEAN difference score, $\bar{d}$: in the previous example $\bar{d}$ = -0.094 kmh
3. Calculate the degrees of freedom: $df = \#\text{ pairs} - 1 = n - 1$, where n is the number of pairs
4. Calculate the standard error of the mean of d.

   variance of $d = s_d^2 = \sum \frac{\left(d_i - \bar{d}\right)^2}{n-1}$

   where

   $SE_{\bar{d}} = \sqrt{\frac{s_{\bar{d}}^2}{n}}$

5. Calculate the test statistic for paired data

$$t = \frac{s^2_{\bar{d}}}{SE_{\bar{d}}}$$

6. Compare to the Critical Value in Appendix Table 2
7. Find the Critical Value = $t_{\alpha(2),df}$

Try as difference instead of paired

Before you answer, take a look at the box plot of the mean difference between the repeat measures of 5K pace for the 15 women. Create a new variable, `raceDiff`, equal to `Race2` minus `Race1`. Then, use the one sample T-test on `raceDiff`. I'll leave you to complete the work (Question 2).

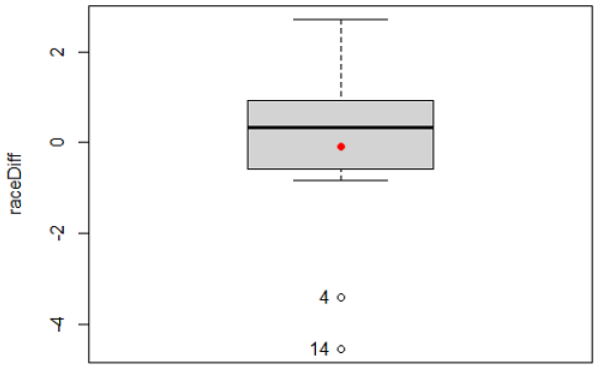


Figure 10.3.5: Box plot of differences. Red dotted lines shows the null hypothesis.

R code

```
t.test(Race1, Race2, paired = TRUE, alternative = "two.sided")
```

R output:

```
Paired t-test

data: Race.1 and Race.2
t = 0.19389, df = 14, p-value = 0.849
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.9491017 1.1377521
sample estimates:
mean of the differences
0.09432517
```

Rcmdr, paired t-test

Rcmdr: Statistics → Means → Paired t-test…

Note: your two groups must be in two different columns (unstacked!) to run this version of the test.

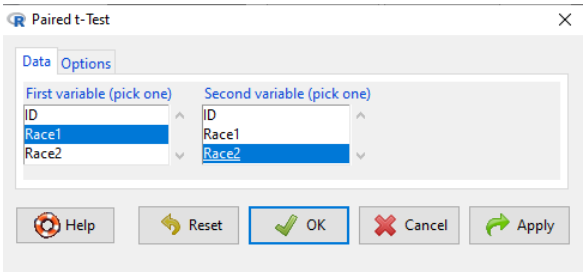


Figure 10.3.6: R Commander Paired t-test menu, Rcmdr version 2.7.

After selecting the variables, set null hypothesis after clicking on Options tab (Fig. 10.3.7).

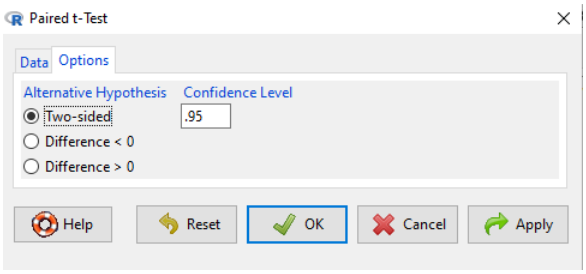


Figure 10.3.7: R Commander Paired t-Test options, select null hypothesis.

Interpret the results.

So, what can we conclude about the null hypothesis? Interpret the 95% CI, the T-test statistic, and the P-value.

Do not ignore sample dependence

What if we ignored the repeated measures design and treated the first and second races as independent? The important concept here is to ask, what would have happened if we had done a two independent sample t-test instead?

Let's run the analysis again, this time incorrectly using the independent sample t-test. We need to manipulate the data set before we do.

Manage your data: Stack the data

This is a good time to share how to Stack data in R. If you look at our active data set, the results of the two trials are in two different columns. In order to run the independent sample t-test we need the data in one column (with a label column).

```
stackExCh10.3 <- stack(example.ch10.3[, c("Race1","Race2")])
names(stackExCh10.3) <- c("obs", "Race")
```

Rcmdr: Data → Active data set → Stack variables in data set…

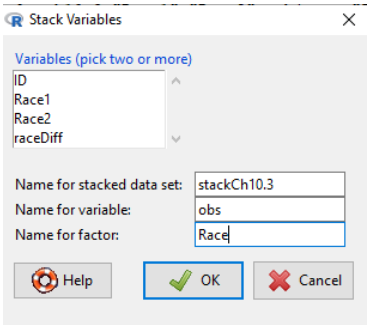


Figure 10.3.8: R Commander: Stack worksheet. Select the two variables Race1 and Race2.

I entered values for name of the new data set, the new variable, and the name for the factor (label) column.

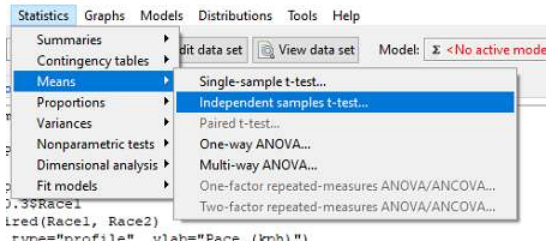


Figure 10.3.9: R Commander, select independent sample t-Test …

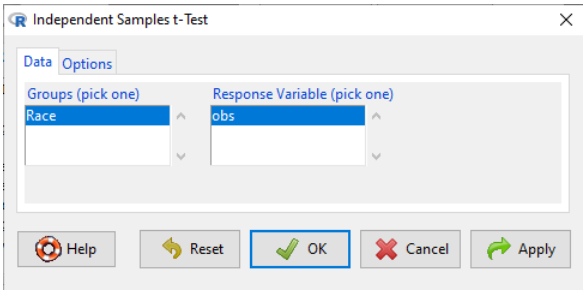


Figure 10.3.10: R Commander, independent sample t-test menu.

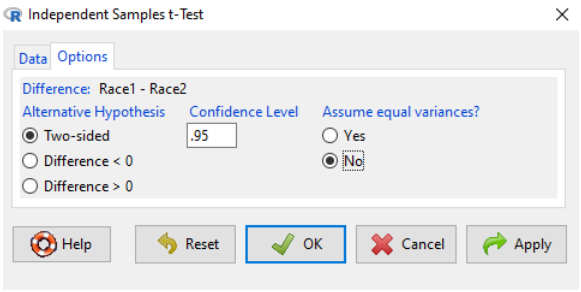


Figure 10.3.11: R Commander, select options for independent sample t-Test (assume equal variance).

Here are the results of the independent sample t-test from R.

```
t.test(obs~Race, alternative='two.sided', conf.level=.95, var.equal=TRUE, + data=stac

Two Sample t-test
data: obs by Race
t = 0.070645, df = 28, p-value = 0.9442
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.640719 2.829369
sample estimates:
mean in group Race1 mean in group Race2
 9.886342 9.792017
```

End R output

In this case, we would have reached the same general conclusion, but the p-values are different. The p-value from the paired t-test was about 0.85 whereas the p-value from the independent sample t-test was higher, nearly 0.95, suggesting little difference between the two trials.

The general conclusion holds this time, that there was no statistically significant difference between the means for first and second trials. However, it won't always work out that way. And besides, if you treated the paired data as independent, you've clearly violated one of the assumptions of the test.

Take a look at the degrees of freedom for the two analyses. By ignoring the pairing of samples we gain twice the number of degrees of freedom … that can't both be right. The way to distinguish between the two is to go back to the experimental units.

Question: What are the sampling units in the case of repeat measures on individuals: the individuals themselves? the pairs of burst speed trials? something else?

it is important to note that the paired t-test is still the best for this situation because it accurately reflects the experiment — individuals were measured twice, therefore the two groups (trial 1 and trial 2) are not independent! Thus, the p-value from the paired t-test correctly reflect our best analyses of the test of the null hypothesis because the correct degrees of freedom were 14 and not 28.

In the case of the independent sample t-test we necessarily make the assumption that the two groups are independent — that is, that they are measured on different sampling units (e.g., different individuals or subjects). In statistical terms, that means that you assume that the correlation between trial results is equal to zero. By incorrectly choosing an independent sample test in these repeated measures cases, I would make two null hypotheses: (1) that the means are the same and (2) that the correlation between repeat measures is zero. The problem? The t-test only evaluates the first hypothesis (means).

Questions

1. Refer to Figure 10.3.5 again, and its related data set. Were runners faster the second year or the first year running the 5k? What about the points in the figure labeled 4 and 14? What was the average difference between first and second races?
2. Complete the test of the null hypothesis of no difference between race 1 and race 2 (`raceDiff`) with the one-sample t-test. Set up a table to compare the test statistic, df, and p-values for results from paired t test, one sample t-test, and independent sample t-test. How do these results compare?
3. I've called the observed value "pace," but runners would know that pace is actually amount of time per kilometer, not the total time over 5k, which is what I called pace.
   - Create a new variable and report average pace for Race1 and Race2.
   - Redo the paired analysis, including box plot, on your new variable.
   - What is the null hypothesis for your new variable?
   - Summarize your results and add to the table you created for question 2.

Data set

```
example.ch10.3 <- read.table(header=TRUE, text = "
ID Race1 Race2
1 15.28 15.61
2 11.22 11.19
3 8.80 9.14
4 8.88 5.46
5 9.81 10.50
6 6.12 5.69
7 8.31 8.71
8 6.26 7.42
9 17.16 16.41
10 16.23 15.82
11 5.90 7.12
12 8.31 10.48
13 5.93 8.64
14 10.54 5.99
15 9.53 8.69
")
```

10.4: Chapter 10 References and Suggested Readings

Andergassen, U., Zebisch, M., Kölbl, A. C., König, A., Heublein, S., Schröder, L., … & Jeschke, U. (2016). Real-time qPCR-based detection of circulating tumor cells from blood samples of adjuvant breast cancer patients: A preliminary study. *Breast Care*, *11*(3), 194-198.

Cohen, J. (1994). The Earth is round (P < 0.05). *American Psychologist* 49:991-1003.

Cowles, M., Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist* 37:553-558.

Driscoll et al. (2000). An introduction to everyday statistics – 2. *Journal of Accident & Emergency Medicine* 17:274-281.

Driscoll, P. (2009). Article 5. An introduction to estimation – 2: from z to t. Emergency Medical Journal 18:65-70.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies—a paradox of statistical practice? BMC Medical Research Methodology 12:78.

Whiteley and Ball (2002). Statistical review 5: Comparison of means. Critical Care6:424-428.

CHAPTER OVERVIEW

11: Power Analysis

Introduction

The **power of a statistical test** is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. Most of us are used to thinking that a hypothesis is either right or it is wrong: a doctor's diagnosis is correct, the patient has the disease, or the patient does not; an experimental result is **objectively true** — i.e., true independent of the observer's subjectivity — or it is false. As we work through a typical science curriculum, we may even take to heart that, unlike mathematicians, scientists don't prove scientific ideas no matter how well supported by evidence. Our acceptance of scientific theories is provisional; if new evidence comes along, we revise and if warranted, we abandon the theory in favor of new explanation. However, even this point does not completely reflect the point we are making from statistical thinking. One of the more challenging concepts for new statistics students to understand is that outcomes of a doctor's diagnosis or of an experiment are associated with probability.

The concept of statistical power helps to relate our ability to confidently conclude one outcome over another. Statistical power depends on

- What Type I error rate we set
- The **effect size** or difference between affected and unaffected groups
- The sample size
- The variability of the subjects

These concepts have all been introduced before, but the idea that even a well designed experiment may lack the capability of detecting "truth" is a new and important topic to add to your growing statistical thinking tool kit.

11.1: What is statistical power?

Introduction

Simply put, the **power of a statistical test** asks the question: do we have enough data to warrant a scientific conclusion, not just a statistical inference? From the NHST approach to statistics, we define two conditions in our analyses: a null hypothesis, H_O, and the alternate hypothesis, H_A. By now you should have a working definition of H_O and H_A (Chapter 8.1). For the two-sided case (Chapter 8.4), H_O would be no statistical difference between two sample means, whereas for H_A, there was a difference between two sample means. Similarly, for one-sided cases, H_O would be one sample mean greater (less) than or equal to the second sample mean, whereas for H_A, one sample mean is less (greater) than the second mean. Together, these two hypotheses cover all possible outcomes of our experiment!

However, when conducting an experiment to test the null hypothesis, four, not two, outcomes are possible with respect to "truth" (Table 11.1.1), which we introduced first in our Risk analysis chapter and again when we introduced statistical inference.

Table 11.1.1. Possible outcomes of an experiment.

		In the population, H_O is really:	
		True	**False**
Inference:	**Fail to reject H_O**	Correct decision $p = 1 - \alpha$	Type II error $p = \beta$
	Reject H_O	Type I error $p = \alpha$	Correct decision $p = 1 - \beta$

We introduced false positives and false negatives in our discussion of risk analysis, and now generalize these concepts to outcomes of any experiment.

> **(1) We do not reject the null hypothesis**, we state that we are 95% $(p = 1 - \alpha)$ confident that we've made the correct decision, **and in fact, that is the true situation ("correct decision")**. As before this is a true positive.

For example, mean acetylsalicylic acid concentration in a sample of 200-mg, brand-name aspirin tablets really is the same as that in generic aspirin.

> **(2) We reject the null hypothesis**, we state there is a 5% chance that we could be wrong $(p = 1 - \beta)$, **and in fact, that is the true situation ("correct decision")**. As before this is a true negative.

For example, ephedrine really does raise heart rates in people who have taken the stimulant compared to those who have taken a placebo.

Two additional possible outcomes to an experiment

The other two possible outcomes are not desirable but may occur because we are making inferences about populations from limited information (we conduct tests on samples) and because of random chance influencing our measure.

> **(3) We do not reject the null hypothesis**, but in fact, **there was a true difference between the two groups and we have therefore committed a Type II error**. As before this is a false negative.
>
> **(4) We reject the null hypothesis**, but in fact, **there was no actual difference between the two groups and we have therefore committed a Type I error**. As before this is a false positive.

What can we do about these two possible undesirable outcomes?

We set our Type I error rate α, and, therefore our Type II error rate β before we conduct any tests, and these error rates cover the possibility that we may incorrectly conclude that the null hypothesis is false (α), or we may incorrectly conclude that the null hypothesis is true (β). When might these events happen?

The ***power of a statistical test*** is the probability that the test will reject the null hypothesis when the alternative hypothesis is true.

A type I error is committed, by chance alone, when our sample is accidentally obtained from the tail of the distribution, thus our sample appears to be different from the population… Below, we have a possible case that, by chance alone, we could be getting all of our subjects from one end of the distribution (Fig. 11.1.1).

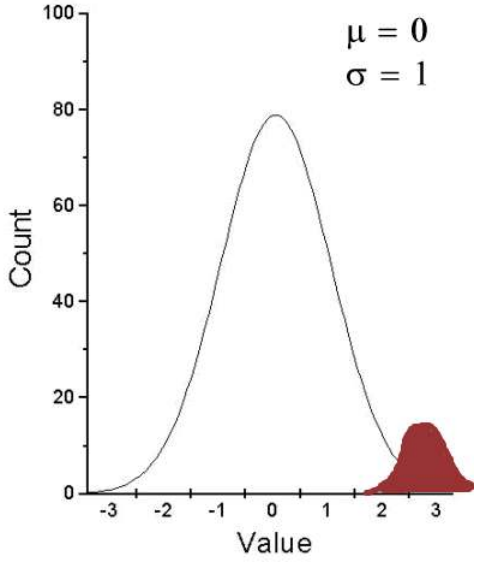


Figure 11.1.1: Population sampling from tail of distribution.

We would likely conclude that our sample mean is different (greater) than the population mean.

A type II error is committed, by chance alone, when our sample is between two different population distributions. The implication for our study is drawn in Figure 11.1.2 Instead of our sampling drawn from one population, we may have drawn between two very different populations.

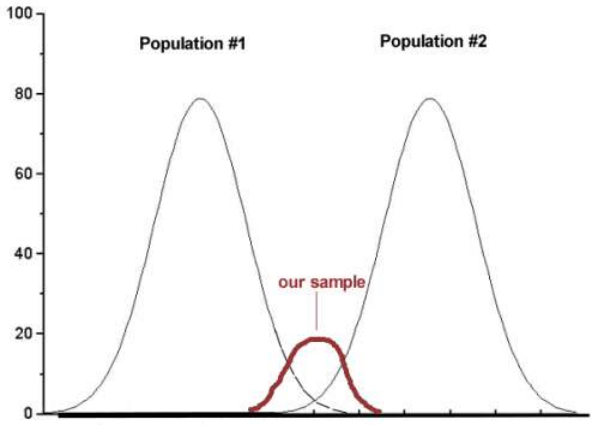


Figure 11.1.2: Without us knowing, our sample may come from the extremes of two separate populations.

How did we end up with "the wrong" sample? Recall from our first discussions about Experimental Design how we distinguished between **random** and **haphazard sampling**. The key concept was that a program of recruitment of subjects (e.g., how to get a sample) must be conducted in such a way that each member of the population has an equal chance of being included in the study. Only then can we be sure that **extrinsic factors** (things that influence our outcome but are not under our control nor studied by us) are spread over all groups, thus canceling out.

Why do we say we are 95% confident in our estimate (or conclusions)?

(1) Because we can never be 100% certain that by chance alone we haven't gotten a **biased sample** (all it takes is a few subjects in some cases to "throw off" our results).

(2) For parametric tests, at least, we assume that we are sampling from a normal population.

Thus, in statistics, we need an additional concept. Not only do we need to know the probability of when we will be wrong (α, β), but we also want to know the probability of when we will be correct when we use a particular statistical test. This latter concept is defined as the power of a test as the probability of correctly rejecting Ho when it is false. Conducting a power analysis before starting the experiment can help answer basic experimental design questions like, how many subjects (experimental units) should my project include? What approximate differences, if any might I expect among the subjects? (Eng 2003; Cohen 1992).

Put another way, power is the likelihood of identifying a significant (important) effect (difference) when one exists. Formally, statistical power is defined as the probability, $p = 1 - \beta$, and it is the subject of this chapter.

Questions

1. True or False. If we reduce Type I error from 5% to 1%, our test of the null hypothesis has more power.
   - Explain your selection to question 1.
2. True or False. Statistical power is linked to Type II error.
   - Explain your selection to question 2.

11.2: Prospective and retrospective power

Introduction

Statistical power is defined as $p = 1 - \beta$. The power of a test is the probability of rejecting a null hypothesis when it is false. There is a relationship between Type I error and Type II error. We want β to be large, BUT β is generally not known when we are performing the statistical test. We do know that α is inversely related to β. The smaller the α value we use to reject the null hypothesis, the MORE likely we will accept a FALSE Null Hypothesis. If we make α very small (one in a billion) then there would be a very HIGH chance of accepting α Null Hypothesis when it is false (β is high). Note that this is the same discussion that we had about sensitivity of an assay test and the specificity of that assay. If we increase sensitivity of the assay such that we approach 100% detection of the true positives will be detected, we necessarily will increase the number of false positives. The BEST way to reduce both Type I and Type II statistical errors is to INCREASE the sample size!

Power of any statistical test (z-test, t-test, one-way ANOVA…) can be determined BEFORE the experiment is done and data are gathered or AFTER an experiment is completed (Cohen 1992). For now, here's our first real taste of experimental design — we can evaluate how we can make an experiment to test a particular hypothesis.

Prospective power analysis

Good experimental design should include considerations of power. The design will determine the size of the effect your experiment will be able to detect and the probability of correctly rejecting the null hypothesis. In large part, this will involve decisions of sample size and ways to reduce error variance. Moreover, one needs to decide ahead of time, just how large of an effect does the experiment need to detect? A one-gram change in body mass before and after a diet treatment is of no concern whatsoever if your study subjects are African elephants, but may be a very large effect for a study of shrews!

Retrospective power analysis

Power analysis can also be done after a test has been conducted and the null hypothesis failed to be rejected. One interpretation that might follow from a retrospective power analysis is that, if the study had low power, the lack of statistical significance could be viewed merely as the result of low sample size. However, as forcefully argued by Hoenig and Heisey (2001) (see also Colegrave and Ruxton 2003), retrospective or post-hoc power tests provide no more information than does the p-value and therefore are redundant. At worse, retrospective power analysis can be misleading as to how well it predicts true power, i.e., biologically meaningful differences between different treatment groups (Zhang et al., 2019).

Prospective power is more than effect size between groups

Large effect size, i.e., large differences between groups, is not necessarily evidence of important biological differences. The concept of power has limits (Hoenig and Heisey 2001, Yuan and Maxwell 2005). On the other hand, small effect sizes can be important differences, especially if the treatment is difficult or expensive. We work through this conclusion with an example, but emphasize here that providing confidence intervals for the effect size (Colegrave and Ruxton 2003). Suppose the null hypothesis was not rejected from a two-sample independent t-test conducted on the test of differences in plant height between two samples of `ohi'a found at different elevations. Was there really no difference or was the sample size in the study simply too small to detect the real and important difference between the samples? By conducting a power analysis, one can determine if a slight increase in sample size would have yielded a statistically significant difference, or it may suggest that the effect size is small enough not to warrant further attention.

R code

Three options for conducting power analysis in R are provided in Section 11.5.

Questions

1. Assuming that a study was done by randomly sampling from a population, and the the primary outcome is found not statistically different with p-value 0.13 between placebo and treatment groups, what can be gained made from a retrospective power analysis?

11.3: Factors influencing statistical power

Introduction

Components of an experimental design that may influence statistical power include:

1. α (probability of committing a Type I error).
   As the probability of a Type I error increases, the probability of a Type II error decreases. Therefore, as α gets bigger, power gets decreases.
2. σ, the variance in the population
   Statistical power decreases as variability increases.
3. Effect size
   The effect size is a measure of the differences between two or more groups considered biologically important; its difficult to have lots of power to detect very small differences.
4. Sample size
   As n increases, power increases.

Size of alpha (Type I error)

When we introduced the idea of Type I error, alpha, it followed a story about the *Challenger*, the space shuttle that disintegrated after the failure of O-rings allowed hot gasses to cross joints in the rockets. If we adopt a Type I error rate of 5% in our rocket designs, then failure is expected at one in twenty launches. That rate clearly is unacceptable, so the logical extension of this thinking would be to decrease the Type I error rate. Unfortunately, this comes at the expense of increasing our risk of Type II error. Thus, decreasing Type I error decreases statistical power.

Variance

If the range of values for individuals in the samples are great, then it should not be surprising that the power to distinguish between sample means from the groups will be low. Conversely, if the variance for a sample is small, then the precision of the estimated sample mean will increase, and, therefore, the power will increase.

Effect size

Effect size deserves some more comments. If we are thinking cause/effect, then we are asking whether or not our independent variable explains a lot of variation in the response (dependent) variable. If there is a strong link between the independent and dependent variable, then the effect size will be large and small numbers of observations will be needed.

There are various ways to estimate effect size, but the simplest is a variation of the t-test (Cohen's d),

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

The formula would be different for ANOVA (involves the mean squares), but you get the idea. By convention, an effect size of about 0.2 would be "small," 0.5 would be "medium," and an effect size greater than 0.8 would be "large."

Sample size

This is the area of course where the experimenter has control. We can choose how many individuals are assigned to treatment groups. The Central Limit Theorem basically states that you can use the normal distribution to predict how likely an individual observation is in relation to a sample mean even if the sample distribution is not normally distributed. The larger the number of individuals in the sample from a population, the more confidence we have about making this assumption. Translation: sample size directly impacts standard error of the calculated statistic. Recall our equation for the standard error of the mean.

Alternatives to Cohen's d

Glass's g

Hedges' g_u

Keselman and colleagues' dj

estimators based on **trimmed mean** and **Winsorized variances**.

Questions

1. The Central limit Theorem can be invoked when we have large samples of observations. In this subchapter we state that increase sample size increases statistical power. Discuss and contrast these two characteristics of large sample size on statistical inference.
2. How many samples are enough? Some statistical textbooks will cite a rule of 30. With respect to factors that affect statistical power, discuss the limitations of adopting such a rule to design an experiment.

11.4: Two-sample effect size

Introduction

An effect size is a measure of the strength of the difference between two samples. The effect size statistic is calculated by subtracting one sample mean from the other and dividing by the pooled standard deviation.

Measures of effect size, **Cohen's d**

$$d = \frac{\bar{X}_1 + \bar{X}_2}{s_{pooled}}$$

where s_{pooled} is the **pooled standard deviation** for the two sample means. An equation for pooled standard deviation was provided in Chapter 3.3, but we'll give it again here.

$$s_{pooled} = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

An alternative version of Cohen's d is available for the t-test test statistic value:

$$d = \frac{2t}{\sqrt{df}}$$

A d of one (1) indicates the effect size is equal to one standard deviation; a d of two (2) indicates the effect size between two sample means is equal to two standard deviations, and so on. Note that effect sizes complement inferential statistics such as p-values.

What makes a large effect size?

Cohen cautiously suggested that values of d

0.2 – small effect size

0.5 – medium effect size

0.8 – large effect size

That is, if the two group means don't differ by much more than 0.2 standard deviations, than the magnitude of the treatment effect is small and unlikely to be biologically important, whereas a $d = 0.8$ or more would indicate a difference of 0.8 standard deviations between the sample means and, thus, likely to be an important treatment effect. Cohen (1992) provided these guidelines based on the following argument. The small effect 0.2 comes from the idea that it is much worse to conclude there is an effect when in fact there is no effect of the treatment rather than the converse (conclude no effect when there is an effect). The ratio of the Type II error (0.2) divided by the Type I error (0.05) gives us the penalty of 4. Similarly, for a moderate effect, $0.5/0.05$ equals 10. Clearly, these are only guidelines (see Lakens 2013).

Examples

The difference in average body size between 6 week old females of two strains of lab mice is 0.4 g (Table 11.4.1), and increases to 1.38 g by 16 weeks (Table 11.4.2).

Table 11.4.1. Average body weights of 6 week old female mice of two different inbred strains.†

Strain	$\bar{X}$	s
C57BL/6J	18.5	0.9
CBA/J	18.1	1.27

†Source: Jackson Laboratories: C57BL/6J; CBA/J

Table 11.4.2. Average body weights of 16 week old female mice of two different inbred strains.†

Strain	$\bar{X}$	s

Strain	$\bar{X}$	s
C57BL/6J	23.9	2.3
CBA/J	25.38	3.76

†Source: Jackson Laboratories: C57BL/6J; CBA/J

The descriptive statistics are based on weights of 360 individuals in each strain (Jackson Labs).

The differences are both statistically significant from a independent t-test, i.e., p-value less than 0.05. I'll show you how to calculate the independent t-test given summary statistics (means, standard deviations), for Table 11.4.1 data, then I will ask you to do this on your own in Questions.

Write an R script, using example data from Table 11.4.1:

```
sdd1 = 0.9
var1 = sdd1^2
sdd2 = 1.27
var2 = sdd2^2
mean1 = 18.5
mean2 = 18.1
n1 = 360
n2 = 360
dff = n1+n2-2
pooledSD <-sqrt((var1+var2)/2)
pooledSEM <-sqrt(var1/n1 + var2/n2); pooledSEM
tdiff<-(mean1-mean2)/pooledSEM; tdiff
pt(tdiff, df=dff, lower.tail=FALSE)
#get two-tailed p-value
2*0.0000006675956
#get cohen's d
2*tdiff/sqrt(dff)
```

Results from the calculations we report (value of the test statistic, degrees of freedom, p-value), and the effect size, then are

```
t = 4.875773, df = 718, p-value = 0.0000006675956
cohen's d = 0.364
```

Now, I'm from the school of "don't reinvent the wheel" or "someone has already solved your problems" (Freeman et al 2008), when it comes to coding problems. And, as you would expect, of course someone has written a function to calculate the t-test given summary statistics. In addition to base R and the `pwr` package (see Chapter 11.5), the package `BSDA` contains several nice functions for power calculations.

To follow this example, install `BSDA`, then run the following code:

```
require(BSDA)
tsum.test(mean1, sdd1, n1, mean2, sdd2, n2, alternative = "two.sided", mu = 0, var.eq
```

R output:

```
Standard Two-Sample t-Test
```

```
data: Summarized x and y
t = 4.8758, df = 718, p-value = 0.000001335
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.2389364 0.5610636
sample estimates:
mean of x mean of y
18.5 18.1
```

Similarly, Cohen's d is available from a package called `effsize` .

Note:

One reason to "re-invent the wheel": I only needed the one function; the BSDA package contains more 330 different objects/functions. A simple way to check how many objects in a package, e.g., BSDA, run

```
ls("package:BSDA")
```

BSDA stands for "Basic Statistics and Data Analysis," and was intended to accompany the 2002 book of the same title by Larry Kitchens.

And of course, if using someone else's code, give proper citation!

Questions

1. We needed an equation to calculate pooled standard error of the mean (`pooledSEM` in the R code). Read the code and write the equation used to calculate the pooled SEM.
2. Calculate the t-test and the effect size for the Table 11.4.1 data, but at three smaller sample sizes. Change 360 to $n_1 = n_2 = 20$, repeat for $n_1 = n_2 = 50$, and finally, repeat for $n_1 = n_2 = 100$. Use your own code, or use the `tsum.test` function from the `BSDA` package.
3. Calculate Cohen's effect size d for each new calculation based on a different sample size.
4. Create a table to report the p-values from the t-tests, the effect size, for each of the four $n_1 = n_2 = (20, 50, 100, 360)$.
5. True or false. The mean difference between sample means remains unaffected by sample size.
6. True or false. The effect size between sample means remains unaffected by sample size.
7. Based on comparisons in your table, what can you conclude about p-value and "statistical significance?" About effect size?
8. Repeat questions 2 – 7 for Table 11.4.2

11.5: Power analysis in R

Introduction

In R, you can estimate the

1. Statistical power — how probable are you to correctly reject the null hypothesis?
2. Sample size — how many samples or observations must we get to have a reasonable chance to correctly reject a null hypothesis?
3. Effect size (minimum difference) — how different are the two samples?

Power analysis is recommended before conducting an experiment, but it is also valuable after an experiment.

Base R and pwr package

In Chapter 11.4 we presented functions from `BSDA` package. R (but not Rcmdr, but see the EZR plugin described below) provides all of the basic power analysis we would need for t-tests, one-way ANOVA, etc. as part of the base installation (Everitt and Hothorn 2007). However, the package `pwr`, provides a more comprehensive package for power analysis. Load and install the R package `pwr`.

For example, to determine the number of samples for an independent sample t-test (two-tailed), the function is `pwr.t.test()`.

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL, type = c("two.sample", "one.sample", "paired")
```

Table 11.5.1. Parameters of the `pwr.t.test()` function.

n	Number of observations (per sample)
d	Effect size
sig.level	Significance level (Type I error probability)
power	Power of test (1 minus Type II error probability)
type	Type of t-test : one-, two-, or paired-samples
alternative	a character string specifying the alternative hypothesis, must be one of “two.sided” (default), “greater” or “less”

The basic question is, how many samples (observations) do I need for each group (k) at Type I error of 5% and type II error of 95%?

To run the test, you fill in all of the “NULL” except the one you wanted solved (i.e., leave n = NULL).

For an effect size of 0.2, Type I error (significance level) of 5%, and 95% power, how many observations per group do we need for our study?

```
pwr.t.test(n = NULL, d =0.2, sig.level = 0.05, power = 0.95, type = c("two.sample"), alternative = c("two.sic
Two-sample t test power calculation
n = 650.6974
d = 0.2
sig.level = 0.05 power = 0.95 alternative = two.sided
```

note: n is number “n” in each group

Wow!

The R `pwr` package is not as convenient as it could be (have to load and run R scripts): for intro level, try one of the many online sites. Here is a website which can help with power analysis based on a variety of situations. The programs are java-coded. http://www.stat.uiowa.edu/~rlenth/Power/index.html

Power analysis with the EZR Rcmdr plugin

The EZR plugin for R Commander provides some facilities to do power analysis (Kanda 2013). First, download and install the RcmdrPlugin.EZR package. The EZR plugin for Rcmdr, `RcmdrPlugin.EZR`, provides an interface to explore power analyses, along with many other statistical functions (Kanda 2013). After loading the plugin to Rcmdr, additional drop down options are added to the menu bar (Fig. 11.5.1).

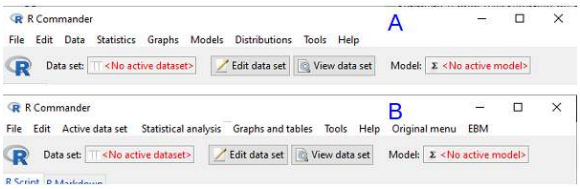


Figure 11.5.1: Screenshot of Rcmdr menu bar with (A) and without (B) the EZR plugin.

I’ll demonstrate use of the plugin, but I recommend that you use `pwr.t.test()` instead. Although EZR uses a drop-down menu system, it has many more functions than we need to solve this simple exercise. Thus, EZR is not any easier to apply for what we need here. That said, off we go.

Worked example

Consider a simple data set: wheel running performance in 24 hours for three strains of mice.

Table 11.5.2. Wheel running performance in 24 hours for three strains of mice.

Mouse strain	Average	Standard deviation	n
AKR	395	169.7	20
CBA	855	77.8	20
C57BL/10	1135	63.6	20

Consider two groups of mice, AKR vs CBA for wheel running performance. How many samples are needed to show a statistical difference for performances between the two groups?

Calculate the pooled standard deviation and calculate a difference between the means (the effect size) for which you wish to say is statistically different. For this example, 132.0 and 460, respectively. With EZR plugin installed and active in R Commander (Fig. 11.5.2), select

Rcmdr: Statistical analysis → Calculate sample size → Calculate sample size for comparison between two means

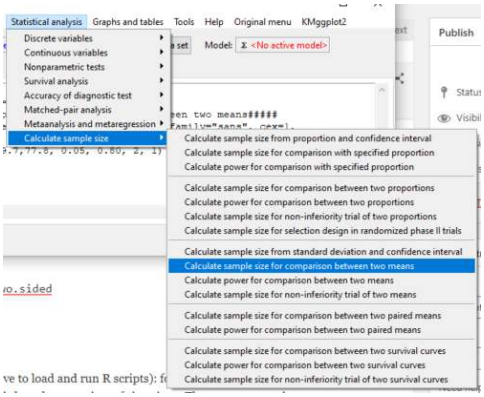


Figure 11.5.2: Screenshot of Rcmdr EZR plugin menu.

Select “Calculate sample size for comparison between two means”, enter the effect size (Difference in means), standard deviation in each group (or a single value for pooled standard deviation), alpha error, power, and sample size ratio.

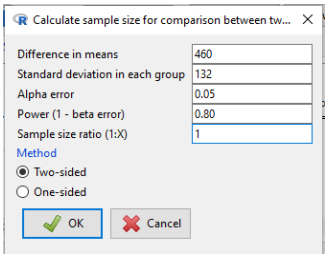


Figure 11.5.2: Screenshot of EZR Menu to obtain sample size for the AKR vs CBA data.

R output from EZR Calculate sample size for comparison between two means:

```
Assumptions
Difference in means 460
Standard deviation 132
Alpha 0.05
two-sided
Power 0.8
N2/N1 1

Required sample size Estimated
N1 2
N2 2
```

Questions

Recall the lizard body mass data set from Chapter 10.1

```
Geckos <- c(3.186, 2.427, 4.031, 1.995)
Anoles <- c(5.515, 5.659, 6.739, 3.184)
```

Enter the data into an R `data.frame` , carry out the independent sample t-test, then

1. Calculate power for the comparison between the two means
2. Calculate sample size needed to achieve 95% power.

11.6: Chapter 11 References and Suggested Readings

Browner, W. S., Newman, T. B. (1987). Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 257:2459-2463.

Cohen, J. (1992). Statistical power analysis. *Current directions in Psychological Science* 1:98-101.

Colegrave, N., and Ruxton, Graeme D. (2003) Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology* 14(3):446-447

Eng, J. (2003). Sample Size Estimation: How Many Individuals Should Be Studied? *Radiology* 227:309-313.

Everitt, B. S., Hothorn, T. (2007) A handbook of statistical analyses using R, 2nd edition. Chapman & Hall/CRC Press.

Freeman, E., Robson, E., Bates, B., & Sierra, K. (2008). *Head first design patterns*. " O'Reilly Media, Inc.".

Hansen, W. B., Collins, L. M. (1994). Seven ways to increase power without increasing N, pp 184-195 in: *Advances in Data Analysis for Prevention Intervention Research*, Collins LM, Seitz LA (eds). NIDA Research Monograph 142.

Hoenig, J. M., Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *American Statistician* 55:19-24.

Kanda, Y. (2013). Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone marrow transplantation* 48:452-458.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863.

Yuan, K.-H., Maxwell, S. (2005). On the Post Hoc Power in Testing Mean Differences. *Journal of Educational and Behavioral Statistics* 30(2):141-167.

Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. (2019). Post hoc power analysis: is it an informative and meaningful analysis?. *General psychiatry*, *32*(4).

CHAPTER OVERVIEW

12: One-way Analysis of Variance

Introduction

We left off with two-group experiments in Chapter 10 where we introduced two-sample tests of the null hypothesis of no difference between the middles for each group (if means, t-tests; if medians, Wilcoxon test).

As review, please revisit what we mean by **independent variables** (statistical jargon for “different treatments, like a placebo vs. aspirin therapy”) and dependent variables (statistical jargon for “**response** or **outcome** of the experiment was recorded as number of living or dead subjects”).

Variables are **independent** in the sense that the values are not related to the experiment’s outcome — we select the levels of the variables. For example, we select to study green vs. red leaves (the variable is “leaf color”, and there are only two levels or states of the variable: green & red). In contrast, we denote the values of the response variable as dependent because the particular values that the variable will take depend on the experiment.

It’s rare that you, as a researcher, would only be interested in comparing two samples or two groups of data for which a treatment has been applied in an experiment or investigation. More often, inferences are drawn on multiple samples (more than two groups) and an experiment involves multiple groups (one or more controls plus one or more experimental treatments).

Previously, we have discussed data sets with only one or two samples or populations (e.g. one- and two-sample t-tests, Mann-Whitney tests). Now we want to extend the discussion of statistics to situations where we may have more than two samples or populations. We introduce the **ANalysis Of VAriance** (**ANOVA**).

Importantly, we will see that one- and two-sample tests are just simple cases of ANOVA. Thus, use of ANOVA should be your preference, even if you have just two groups.

12.1: The need for ANOVA

Introduction

Moving from an experiment with two groups to multiple groups is deceptively simple: we move from one comparison to **multiple comparisons**. Consider an experiment in which we have randomly assigned patients to receive one of three doses of a statin drug (lower cholesterol), including a placebo (e.g., Tobert and Newman 2015). Thus, we have three groups or **levels** of a single treatment **factor** and we'll want to test the null hypothesis that the group (level) means are all equal as opposed to the alternate hypothesis in which one or more of the group means, e.g., group A, group B, group C, are different.

$$H_0 : \bar{X}_A = \bar{X}_B = \bar{X}_C$$

The correct procedure is to analyze multiple levels of a single treatment with a one-way analysis of variance followed by a suitable **post-hoc** ("after this") test. Two common post-hoc tests are **Tukey's range test** (aka Tukey's HSD [honestly significant difference] test), which is for all **pairwise comparisons**, or the **Dunnett's test**, which compares groups against the control group. Post-hoc tests are discussed in Chapter 12.6.

Thus, the **family-wise** (aka **experiment-wise**) error rate for multiple comparisons is kept at 5%, and each **individual-wise comparison** is compared against a more strict (i.e., smaller **Type I error rate**). Put another way, the family-wise error rate is the chance of a number of false positives: making a mistake when we consider many tests simultaneously. The simplest correction for the individual-wise error rate is the **Bonferroni correction**: test each individual comparison at Type I error equal to α/C, where C is the number of comparisons.

> **Note:**
>
> To get the number of "pairwise" comparisons (C), let
>
> $$C = \frac{k(k-1)}{2}$$

For our three group experiment, how many pairwise comparisons can be tested? Therefore, $k = 3$, and we have

$$C = \frac{3(3-1)}{2} = 3$$

Thus, for our three groups, A, B, and C, there are three possible pairwise comparisons.

$$H_1 : \bar{X}_A = \bar{X}_B$$
$$H_2 : \bar{X}_A = \bar{X}_C$$
$$H_3 : \bar{X}_B = \bar{X}_C$$

How many pairwise comparisons for a four-group experiment? Check your work, you should get $C = 6$.

The multiple comparison problem

Let's say that we're stubborn. We could do many single two-sample t-tests — certainly, your statistical software won't stop you — but this is a situation that calls for statistical reasoning. Here's why we should not: we will increase the probability of rejecting a null hypothesis when the null hypothesis is true (e.g., discussion in Jafari and Ansari-Pour 2019). That is, the chance we will commit a Type I error increases if we do not account for the lack of independence in these sets of pairwise tests evaluated by t-tests. This is **multiplicity** or the **multiple comparison problem**.

Review: when we perform a two-sample t-test we are willing to reject a true null hypothesis 5% of the time. This is what is meant by setting the critical probability value (alpha) = 0.05. By "willing" we mean that we know that our conclusions could be wrong because we are working with samples, not the entire population. (Of course at the time, we have no way of actually knowing WHEN we are wrong, but we do want to know how likely we could be wrong!) However, if we compare three population means we have three separate null hypotheses.

H_O: one or more of the means are different.

But if we conduct these as separate independent sample t-tests, then we are implicitly making the following null hypothesis statements:

$$H_1 : \bar{X}_O = \bar{X}_B$$
$$H_2 : \bar{X}_O = \bar{X}_C$$
$$H_3 : \bar{X}_O = \bar{X}_C$$

Thus, we have a 5% chance of being wrong for the first hypothesis and/or a 5% chance of being wrong for the second hypothesis and/or a 5% chance of being wrong for the third hypotheses. The chance that we will be wrong for at least one of these hypotheses must now be greater than 5%.

For three separate hypotheses there is a 14% chance of being wrong when we have the probability value for each individual t-test set at $\alpha = 0.05$. How did we get this result? The point is that these tests are not independent, they are done on the same data set; therefore, you can't simply apply the multiplication probability rule.

Here's how to figure this: for the set of three hypotheses, the probability of incorrectly rejecting at least one of the null hypotheses is $1-(1-\alpha)^3 = 1-0.957 = 0.143$

So, for three t-tests on the same experiment, the Type I error for the overall tests (experiment-wise) is actually 14%, not 5%. It gets worse as the number of combinations (groups and therefore hypotheses) increases. For four groups, Type I error is actually $1-(1-\alpha)^4 = 1-0.815 = 0.185$.

That's 18.5%, not 5%.

If we have just five populations means to compare, the probability of rejecting a null hypothesis when it is true climbs to 60%! How did this happen? The probability of correctly rejecting all of them is now $(1-\alpha)^5 = 0.774$

So, the probability of incorrectly rejecting one test (Type I error) is now $1-(1-\alpha)^5 = 1-0.774 = 0.226 = 22.6\%$ instead of the 5% we think we are testing.

This is the key argument for why you must use ANOVA to analyze multiple samples instead of a combination of t-tests!! ANOVA guarantees that the overall error rate is the specified 5%.

Why is the Type I error not 5% for each test? Because we conducted ONE experiment, we can conduct only ONE test (we could be right, we could be wrong 5% of the time). If we conduct the experiment over again, on new subjects, each time resulting in new and therefore independent data sets, then Type I error = 5% for each of these independent experiments.

Now, I hope I have introduced you to the issue of Type I error at the level of a single comparison and the idea of an experiment, holding Type I error-rates at 5% across all hypotheses to be evaluated in an experiment. You may wonder why anyone would make this mistake now. Actually, people make this "mistake" all the time and in some fields like evaluating gene expression for microarray data, this error was the norm, not the exception (see, for example, discussion of this in Jeanmougin et al 2010).

To conclude, if one does multiple tests on the same experiment, whether it is t-tests or some other test for that matter, then our subsequent tests are related. This is what we mean by independence in statistics — and there are many ways that nonindependence may occur in experimental research. For example, we introduced the concept of **pseudoreplication**, when observations are treated as if they are independent, but they are not (see Chapter 5.2). The "multiple comparison problem" specifically refers to the lack of independence when all the data set from a single experiment is parsed into lots of separate tests. Philosophically, there must be a logical penalty — and that is reflected in the increase in Type I error.

Clearly something must be done about this!

ANOVA is a solution

One possible solution for getting the correct experiment-wise error rate: adjust for differences in probability for multiple comparisons with the t-test. We used post-hoc tests presented above: you could evaluate the tests after accounting for the change in Type I error. This is what is done in many cases. For example, in genomics. In the early days of gene expression profiling by microarray, it was common to see researchers conduct t-tests for each gene. Since microarrays can have thousands of genes represented on the chip, then these researchers were conducting thousands of t-tests, arranging the t-tests by P-value and counting the number of p-values less than 5% and declaring that the differences were statistically significant.

This error didn't stand long, and there are now many options available to researchers to handle the "multiple comparisons" problem (some probably better than others, research on this very much an ongoing endeavor in biostatistics). The Bonferroni correction was

an available solution, largely replaced by the **Holm** (aka **Holm-Bonferroni**) **method** (Holm 1979). Recall that the Bonferroni correction judged individual p-values statistically significant only if they were less than α/C, where, again, α is the family-wise error rate (e.g., $\alpha = 0.05$) and C was the number of comparisons. The Holm method orders the C p-values from lowest to highest rank. The method then evaluates lowest p-value, if less than α/C, then reject hypothesis for that comparison. Proceed to next p-value, if less than α/C, then reject hypothesis for that comparison, and so on until no more comparisons are less than α/C.

A MUCH better alternative is to perform a single analysis that takes the multiple-comparisons problem into account: single-factor ANOVA, also called the one-way ANOVA, plus the **post-hoc tests** with error correction. We introduce one-way ANOVA in the next section. Post-hoc tests are discussed in Chapter 12.6.

Questions

1. You should be able to define and distinguish how Bonferoni correction, Dunnett's test, and Tukey's test methods protect against inflation of Type I error.
2. What will be the experiment-wise error rate for an experiment in which there are only two treatment groups?
3. Experiment-wise error rate may also be called _____ error rate.
4. List and compare the three described posthoc approaches to correct for multiple comparison problem,
5. Glycophosphate-tolerant soy bean is the number one GMO (genetically modified organism) crop plant worldwide. Glycophosphate is the chief active ingredient in Roundup, the most widely used herbicide. A recent paper examined "food quality" of the nutrient and elemental composition of plants drawn from fields which grow soy by organic methods (no herbicides or pesticides) and GMO plants subject to herbicides and pesticides. A total of 28 individual t-tests were used to compare the treatment groups for different levels of nutrients and elements (e.g., vitamins, amino acids, etc.,); the authors concluded that 10 of these t-tests were statistically significant at Type I error rate of 5%. Discuss the approach to statistical inference by the authors of this report; include correct use of the terms experiment-wise and individual-wise in your response and suggest an alternative testing approach if it is appropriate in your view.
6. In a comparative study about resting metabolic rate for eleven species of mammals, how many pairwise species comparisons can the study test?
7. In Chapter 4.2 we introduced a data set from an experiment. The experiment looked at DNA damage quantified by measuring qualities in a Comet Assay including the Tail length, the percent of DNA in the tail, and olive moment. The data set is copied to end of this page. In the next chapter I'll ask you to conduct the ANOVA on this experiment. For now, answer the following questions.
   a. What is the response variable?
   b. Explain why there is only one response variable.
   c. How many treatment variables are there?
   d. Why is this an ANOVA problem? Include as part of your explanation a statement of the null hypothesis.

Data used in this page

comet assay dataset

Data set, comet assay

Table 12.1.1. Comet assay data.

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	10	9.7732	2.1501
Copper-Hazel	6	4.8381	0.9676
Copper-Hazel	6	3.981	0.836
Copper-Hazel	16	12.0911	2.9019
Copper-Hazel	20	15.3543	3.9921
Copper-Hazel	33	33.5207	10.7266
Copper-Hazel	13	13.0936	2.8806
Copper-Hazel	17	26.8697	4.5679

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	30	53.8844	10.238
Copper-Hazel	19	14.983	3.7458
Copper	11	10.5293	2.1059
Copper	13	12.5298	2.506
Copper	27	38.7357	6.9724
Copper	10	10.0238	1.9045
Copper	12	12.8428	2.5686
Copper	22	32.9746	5.2759
Copper	14	13.7666	2.6157
Copper	15	18.2663	3.8359
Copper	7	10.2393	1.9455
Copper	29	22.6612	7.9314
Hazel	8	5.6897	1.3086
Hazel	15	23.3931	2.8072
Hazel	5	2.7021	0.5674
Hazel	16	22.519	3.1527
Hazel	3	1.9354	0.271
Hazel	10	5.6947	1.3098
Hazel	2	1.4199	0.2272
Hazel	20	29.9353	4.4903
Hazel	6	3.357	0.6714
Hazel	3	1.2528	0.2506

Rat lung cells treated with Hazel tea extract and exposed to copper metal. Tail refers to length of the comet tail, TailPercent is percent DNA damage in tail, and Olive moment refers to Olive's (1990), defined as the fraction of DNA in the tail multiplied by tail length.

12.2: One-way ANOVA

Introduction

In ANalysis Of VAriance, or **ANOVA** for short, we likewise have tools to test the null hypothesis of no difference between between categorical **independent variables** — often called **factors** when there's just a few **levels** to keep track of — and a single, dependent response variable. But now, the response variable is quantitative, not qualitative like the χ^2 tests.

Analysis of variance, ANOVA, is such an important statistical test in biology that we will take the time to "build it" from scratch. We begin by reminding you of where you've been with the material.

We already saw an example of this null hypothesis. When there's only one factor (but with two or more levels), we call the **analysis of means** and "one-way ANOVA." In the independent sample t-test, we tested whether two groups had the same mean. We made the connection between the confidence interval of the difference between the sample means and whether or not it includes zero (i.e., no difference between the means). In ANOVA, we extend this idea to a test of whether two or more groups have the same mean. In fact, if you perform an ANOVA on only two groups, you will get the exact same answer as the independent two-sample t-test, although they use different distributions of critical values (t for the t-test, F for the ANOVA — a nice little fact for you, if you square the t-test statistic, you'll get the F-test statistic: $t^2 = F$).

Let's say we have an experiment where we've looked at the effect of different three different calorie levels on weight change in middle-aged men.

I've created a simulated dataset which we will use in our ANOVA discussions. The data set is available at the end of this page (scroll down or click here).

We might graph the mean weight changes $(\pm SEM)$. Below are two possible outcomes of our experiment (Fig. 12.2.1).

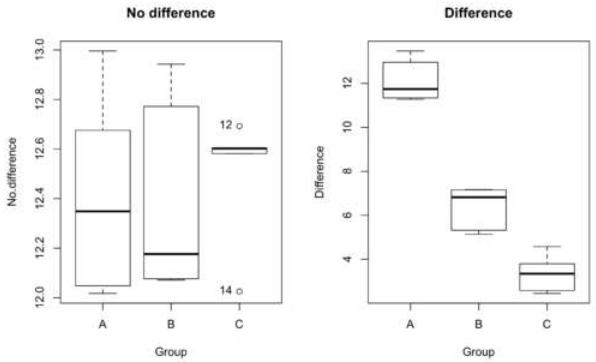


Figure 12.2.1: Hypothetical results of an experiment, as box plots. Left, no difference among groups; Right, large differences among groups.

As statisticians and scientists, we would first calculate an overall or **grand mean** for the entire sample of observations; we know this as the sample mean whose symbol is $\bar{X}$. But this overall mean is made up of the average of the sample means. If the null hypothesis is true, then all of the sample means all estimate the overall mean. Put another way, the null hypothesis being true means that being a member of a treatment group doesn't matter, i.e., there is no systematic effect, and all differences among subjects are due to random chance.

$$H_O : \bar{X} = \bar{X}_A = \bar{X}_B = \bar{X}_C$$

The hypotheses among are three groups or treatment levels then are:

$$H_O : \bar{X}_A = \bar{X}_B = \bar{X}_C$$

The null hypothesis is that there are no differences among the group means. And the alternative hypotheses include any (or all) of the following possibilities:

$$H_O : \bar{X}_A \neq \bar{X}_B = \bar{X}_C$$

or maybe

$$H_O : \bar{X}_A = \bar{X}_B \neq \bar{X}_C$$

or… have we covered all possible alternate outcomes among three groups?

In either case, we could use one-way ANOVA to test for "statistically significant differences."

Three important terms you'll need for one-way ANOVA

FACTOR: We have one factor of interest. For example, a factor of interest might be

- Diet fed to hypertensive subjects (men and women)
- Distribution of coral reef sea cucumber species in archipelagos
- Antibiotic drug therapy for adolescents with *Acne vulgaris* (see Webster 2002 for review).

LEVELS: We can have multiple levels (2 or more) within the single factor. Some examples of levels for the Factors listed:

- Three diets (DASH, diet rich in fruits & vegetables, control diet)
- Five archipelagos (Hawaiian Islands, Line Islands, Marshal Islands, Bonin Islands, and Ryukyu Islands)
- Five antibiotics (ciprofloxacin, cotrimoxazole, erythromycin, doxycycline, minocycline).

RESPONSE: There is one **outcome** or **measurement** variable. This variable must be quantitative (i.e., on the ratio or interval scale). Continuing our examples then

- Reduction in systolic pressure
- Numbers of individual sea cucumbers in a plot
- Number of microcomedo†.

†A comedo is a clogged pore in the skin; a microcomedo refers to the small plug. Yes, I had to look that up, too.

The response variable can be just about anything we can measure, but because ANOVA is a parametric test, the response variable must be Normally Distributed!

Note on experimental design

As we discuss ANOVA, keep in mind we are talking about analyzing results from an experiment. Understanding statistics, and in particular ANOVA, informs how to plan an experiment. The basic experimental design is called a **completely randomized experimental design**, or **CDR**, where treatments are assigned to **experimental units** at random.

In this experimental design, subjects (experimental units) must be randomly assigned to each of these levels of the factor. That is, each individual should have had the same probability of being found in any one of the levels of the factor. The design is complete because randomization is conducted for all levels, all factors.

Thinking about how you would describe an experiment with three levels of some treatment, we would have the following:

Table 12.2.1. Summary statistics of three levels for some ratio-scale response variable.

Level 1	$\bar{X}_1$ sample mean$_1$	s_1 sample standard deviation$_1$
Level 2	$\bar{X}_2$ sample mean$_2$	s_2 sample standard deviation$_2$
Level 3	$\bar{X}_3$ sample mean$_3$	s_3 sample standard deviation$_3$

Note:

This table is the basis for creating the box plots in Figure 12.2.1.

ANOVA sources of variation

ANOVA works by partitioning **total variability** about the means (the grand mean, the group means). We will discuss the multiple samples and how the ANOVA works in terms of the sources of variation. There are two "sources" of variation that can occur:

- **Within Group Variation**
- **Among Groups Variation**

So let's look first at the variability within groups, also called the **Within Group Variation**.

Consider an experiment to see if DASH diet reduces systolic blood pressure in USA middle-aged men and women with hypertension (Moore et al 2001). After eight weeks we have

	Control Diet (n=25)	Fruit/Vegetable Diet (n=24)	DASH Diet (n=23)
Decrease in SBP, mmHg	0.6	3.8	11.8

We get the corrected **sum of squares**, SS, for within groups:

$$\text{within-groups } SS = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right)^2 \right]$$

and the degrees of freedom, DF, for within groups:

$$\text{within-groups } DF = \sum_{i=1}^{k} (n_i - 1) = N - k$$

where i is the identity of the groups, X_{ij} is the individual observations within group i, $\bar{X}_i$ is the group i mean, n_i is the sample size within each group, N is the total sample size of all subjects for all groups, and k is the number of groups.

Importantly, this value is also referred to as "**error sums of squares**" in ANOVA. Its importance is as follows — In our example, the within-group variability would be zero if and only if all subjects within the same diet had the same reduction in systolic blood pressure. This is hardly ever the case of course in a real experiment. Because there are almost always some other unknown factors or measurement error that affect the response variable, there will be some unknown variation among individuals who received the same treatment (within the same group). Thus, the **error variance** will generally be larger than zero.

The first point to consider: your ANOVA will never result in statistical differences among the groups if the error variance is greater than the second type of variability, the variability between groups.

The second type of variability in ANOVA is that due to the groups or treatments. For example, if the response variable being measured was body weight, individuals given a calorie-restricted diet will lose some weight; individuals allowed to eat a calorie-rich diet likely will gain weight, therefore there will be variability (a difference) due to the treatment. So we can calculate the variability among groups. We get the corrected sum of squares for among groups:

$$\text{among-groups } SS = \sum_{i=1}^{k} n_i \left(X_i - \bar{X} \right)^2$$

and the degrees of freedom for among groups:

$$\text{among-groups } DF = k - 1$$

where i is the identity of the groups, $\bar{X}$ is the grand mean as defined in Measures of Central Tendency (Chapter 3.1), $\bar{X}_i$ is the group i mean, n_i is the sample size within each group, N is the total sample size of all subjects for all groups, and k is the number of groups.

The sums of squares here is simply subtracting the mean of each population from the overall mean.

- If the Factor is not important in explaining the variation among individuals then all the population means will be similar and the sums of squares among populations would be small.
- If the Factor is important in explaining some of the variation among the individuals then all the population means will NOT be the same and the sums of squares among populations would be large.

Finally, we can identify the total variation in the entire experiment. We have the total sum of squares.

$$\text{Total } SS = \text{among-groups } SS + \text{within-groups } SS$$

Thus, the insight of ANOVA is that variation in the dataset may be attributed to a portion explained by differences among the groups and differences among individual observations within each group. The inference comes from recognizing that if the among group effect is greater than the within group effect, then there will be a difference due to the treatment levels.

Mean squares

To decide whether the variation associated with the among group differences are greater than the within group variation, we calculate ratios of the sums of squares. These are called **Mean Squares** or MS for short. The ratio of the Mean Squares is called F, the test statistic for ANOVA.

For the one-way ANOVA we will have two Mean Squares and one F, tested with degrees of freedom for both the numerator MS_{groups} and the denominator MS_{error}.

The **Mean Square for (among) groups** is

$$MS_{group} = \frac{\text{among-groups } SS}{\text{among-groups } SS}$$

The **Mean Square for error** is

$$MS_{error} = \frac{\text{within-groups } SS}{\text{within-groups } SS}$$

And finally, the value for F, the test statistic for ANOVA, is

$$F = \frac{MS_{groups}}{MS_{error}}$$

Worked example with R

A factor with three levels, A, B, and C

```
group <- c("A", "A", "A", "B", "B", "B", "C", "C", "C")
```

and their responses, simulated

```
response <- c(10.8, 11.8, 12.8, 6.5, 7, 8, 3.8, 2.8, 3)
```

We create a data frame

```
all <- data.frame(group, response)
```

Of course, you could place the data into a worksheet and then import the worksheet into R. Regardless, we now have our dataset.

Now, call the ANOVA function, `aov`, and assign the results to an object (e.g., Model.1)

```
Model.1 <- aov(response ~ group, data=all)
```

Now, visualize the **ANOVA table**

```
summary(Model.1)
```

and the output from R, the ANOVA table, is shown below:

Table 12.2.2 *Output from* `aov()` *command, the ANOVA table, for the "Difference" outcome variable.*

```
            Df   Sum Sq    Mean Sq   F value       Pr(>F)
group        2   111.16      55.58     89.49    0.0000341 ***
Residuals    6     3.73       0.62
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's take the ANOVA table one row at a time.

- The first row has subject headers defining the columns.
- The second row of the table "groups" contains statistics due to group, and provides the comparisons among groups.
- The third row, "Residuals" is the error, or the differences within groups.

Moving across the columns, then, for each row, we have in turn,

- the degrees of freedom (there were 3 groups, therefore 2 DF for group),
- the Sums of Squares, the Mean Squares,
- the value of F, and finally,
- the P-value.

R provides a helpful guide on the last line of the ANOVA summary table, the " `Signif[icance] codes` ," which highlights the magnitude of the P-value.

What to report? ANOVA problems can be much more complicated than the simple one-way ANOVA introduced here. For complex ANOVA problems, report the ANOVA table itself! But for the one-way ANOVA it would be sufficient to report the **test statistic**, the **degrees of freedom**, and the **p-value**, as we have in previous chapters (e.g., t-test, chi-square, etc.). Thus, we would report:

F = 89.49, df = 2 and 6, p = 0.0000341

where `F = 89.49` is the test statistic, `df = 2` (degrees of freedom for the among group mean square) and 6 (degrees of freedom for the within group mean square), and `p = 0.0000341` is the p-value.

In Rcmdr, the appropriate command for the one-way ANOVA is simply

Rcmdr: Statistics → Means → One-way ANOVA…

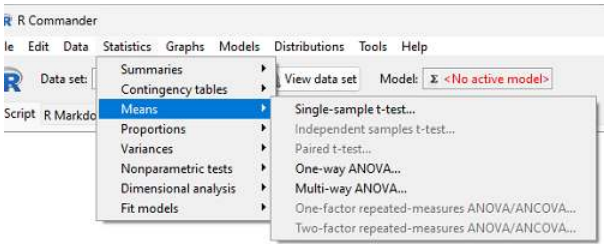


Figure 12.2.2: Screenshot in Rcmdr to select one-way ANOVA.

which brings up a simple dialog. R Commander anticipates factor (Groups) and Response variable. Optional, choose **Pairwise comparisons** of means for **post-hoc test** (**Tukey's)** and, if you do not want to assume equal variances (see Chapter 13), select **Welch F-test**.

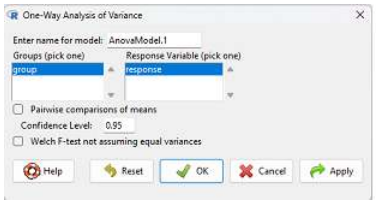


Figure 12.2.3: Screenshot of selecting one-way ANOVA options in Rcmdr.

Questions

1. Review the example ANOVA Table (Table 12.2.2 and confirm the following

- How many levels of the treatment were there?
- How many sampling units were there?
- Confirm the calculation of MS_{group} and MS_{error} using the formulas contained in the text.
- Confirm the calculation of F using the formula contained in the text.
- The degrees of freedom for the F statistic in this example were 2 and 6 ($F_{2,6}$). Assuming a two-tailed test with Type I error rate of 5%, what is the critical value of the F distribution (see Appendix A.5)?

2. Repeat the one-way ANOVA using the simulated data, but this time, calculate the ANOVA problem for the " `No.difference` " response variable.

3. Leaf lengths from three strains of *Arabidopsis thaliana* plants grown in common garden are shown in Fig. 12.2.4 Data are provided for you in the following R script.

```
arabid <- c("wt","wt","wt","AS1","AS1","AS1","AS2","AS2","AS2")
leaf <- c(4.909,5.736,5.108,6.956,5.809,6.888,4.768,4.209,4.065)
leaves <- data.frame(arabid,leaf)
```

- Write out a null and alternative hypotheses
- Conduct a test of the null hypothesis by one-way ANOVA
- Report the value of the test statistic, the degrees of freedom, and the P-value
- Do you accept or reject the null hypothesis? Explain your choice.

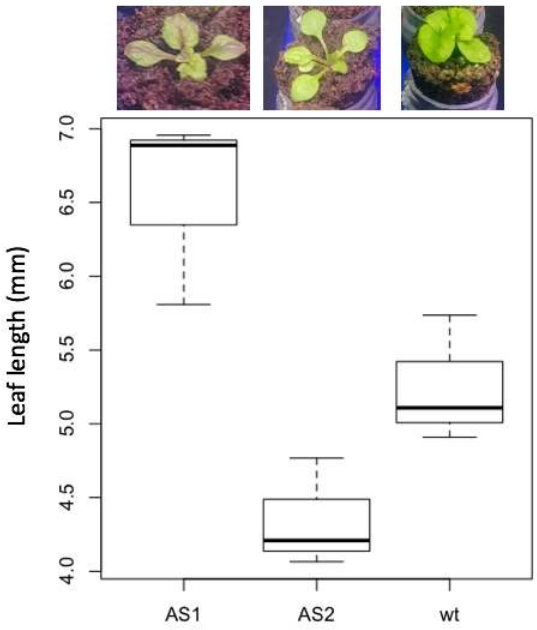


Figure 12.2.4: Box plot of lengths of leaves on one 10-day old plant from each of three strains of *Arabidopsis thaliana*.

4. Return to your answer to question 7 from Chapter 12.1 and review your answer and modify as appropriate to correct your language to that presented here about factors and levels.

5. Conduct the one-way ANOVA test on the Comet assay data presented in question 7 from Chapter 12.1. Obtain the ANOVA table and report the value of the test statistics, degrees of freedom, and p-value.
a. Based on the ANOVA results, do you accept or reject the null hypothesis? Explain your choice.

Data used in this page

Difference, no difference

Difference or No Difference

Table 12.2.3. Difference or no difference.

Group	No.difference	Difference
A	12.04822	11.336161
A	12.67584	13.476142

Group	No.difference	Difference
A	12.99568	12.96121
A	12.01745	11.746712
A	12.34854	11.275492
B	12.17643	7.167262
B	12.77201	5.136788
B	12.07137	6.820242
B	12.94258	5.318743
B	12.0767	7.153992
C	12.58212	3.344218
C	12.69263	3.792337
C	12.60226	2.444438
C	12.02534	2.576014
C	12.6042	4.575672

Table 12.2.1 consists of simulated data.

Comet assay, antioxidant properties of tea

Data presented in Chapter 12.1

12.3: Fixed effects, random effects, and ICC

Introduction

Within discussions of one-way ANOVA models, the distinction between two general classes of models needs to be made clear by the researcher. The distinction lies in how the levels of the factor are selected. If the researcher selects the levels, then the model is a **Fixed Effects Model**, also called a **Model I ANOVA**. On the other hand, if the levels of the factor were selected by random sampling from all possible levels of the factor, then the model is a **Random Effects Model**, also called a **Model II ANOVA**.

Here's an example to help the distinction. Consider an experiment to see if over-the-counter painkillers are as good as prescription pain relievers at reducing numbers of migraines over a six-week period. The researcher selects Tylenol®, Advil®, Bayer® Aspirin, and Sumatriptan (Imitrex®), the latter an example of a medicine only available by prescription. This is clearly an example of fixed effects; the researcher selected the particular medicines for use.

Random effects, in contrast, implies that the researcher draws up a list of all over-the-counter pain relievers and draws at random three medicines; the researcher would also randomly select from a list of all available prescription medicines.

Fixed effects are probably the more common experimental approaches. To be complete, there is a third class of ANOVA called a Mixed Model or Model III ANOVA, but this type of model only applies to multidimensional ANOVA (e.g., two-way ANOVA or higher), and we reserve our discussion of the Model III until we discuss multidimensional ANOVA (Table 12.3.1).

Table 12.3.1. ANOVA models.

ANOVA model	**Treatments are**
I	Fixed effects
II	Random effects
III	Mixed, both fixed & random effects

Although the calculations for the one-way ANOVA under Model I or Model II are the same, the interpretation of the statistical significance is different between the two.

In Model I ANOVA, any statistical difference applies to the differences among the levels selected, but cannot be generalized back to the population. In contrast, statistical significance of the Factor variable in Model II ANOVA cannot be interpreted as specific differences among the levels of the treatment factor, but instead, apply to the population of levels of the factor. In short, Model I ANOVA results apply only to the study, whereas Model II ANOVA results may be interpreted as general effects, applicable to the population.

This distinction between fixed effects and random effects can be confusing, but it has broad implications for how we interpret our results in the short-term. This conceptual distinction between how the levels of the factor are selected also has general implications for our ability to acquire generalizable knowledge by meta-analysis techniques (Hunter and Schmidt 2000). Often we wish to generalize our results: we can do so only if the levels of the factor were randomly selected in the first place from all possible levels of the factor. In reality, this may not often be the case. It is not difficult to find examples in the published literature in which the experimental design is clearly fixed effects (i.e., the researcher selected the treatment levels for a reason), and yet in the discussion of the statistical results, the researcher will lapse into generalizations.

Random Effects Models and Intraclass Correlation Coefficient (ICC)

Model II ANOVA is common in settings in which individuals are measured more than once. For example, in behavioral science or in sports science, subjects are typically measured for the response variable more than once over a course of several trials. Another common setting of Model II ANOVA is where more than one raters are judging an event or even a science project. In all of these cases what we are asking is about whether or not the subjects are consistent, in other words, we are asking about the precision of the instrument or measure.

In the assessment of learning by students, for example, different approaches may be tried and the instructor may wish to investigate whether the interventions can explain changes in test scores. There are an enormous number of articles on reliability measures in the social sciences and you should be aware of a classical paper on reliability by Shrout and Fleiss (1979) (see also McGraw and Wong, 1996). Both the ICC and the product moment correlation, r, which we will introduce in Chapter 16, are measures of strength

of linear association between two ratio scale variables (Jinyuan et al 2016). But ICC is more appropriate for association between repeat measures of the same thing, e.g., repeat measures of running speed. In contrast, the product moment correlation can be used to describe association between any two variables, e.g., between repeat measures of running speed, but also between, say, running speed and maximum jumping height. The concept of **repeatability** of individual behavior or other characteristics is also a common theme in genetics, and so you should not be surprised to learn that the concept actually traces to RA Fisher and his invention of ANOVA and, like in the sociology literature, there are many papers on the use and interpretation of repeatability in the evolutionary biology literature (e.g., Lessels and Boag 1987; Boake 1989; Dohm 2002; Wolak et al 2012).

There are many ways to analyze these kinds of data, but a good way is to treat this problem as a one-way ANOVA with Random Effects. Thus, the Random Effects model permits the partitioning of the variation in the study into two portions: the amount that is due to differences among the subjects or judges or intervention versus the amount that is due to variation within the subjects themselves. The Factor is the Subjects and the levels of the factor are how ever many subjects are measured twice or more for the response variable.

If the subjects performance is repeatable, then the Mean Square Between (Among) Subjects, MS_B, component will be greater than the Mean Square Error component, MS_W, of the model. There are many measures of repeatability or reliability, but the intraclass correlation coefficient, or ICC, is one of the most common. The ICC may be calculated from the Mean Squares gathered from a Random Effects one-way ANOVA. ICC can take any value between zero and one.

$$ICC = \frac{s_B^2}{s_B^2 - s_W^2}$$

where $s_B^2 = MS_B - \frac{MS_W}{k}$ and $s_W^2 = MS_W$

B and W refer, respectively, to the among group (between- or among-groups mean square) and the within group components of variation (error mean square), from the ANOVA. MS refers to the Mean Squares, and k is the number of repeat measures for each experimental unit. In this formulation k is assumed to be the same for each subject.

By example, when a collection of sprinters run a race, if they ran it again, would the outcome be the same, or at least predictable? If the race is run over and over again and the runners cross the finish lines at different times each race, then much of the variation in performance times will be due to race differences largely independent of any performance abilities of the runners themselves and the Mean Square Error term will be large and the Between subjects Mean Square will be small. In contrast, if the race order is preserved race after race: Jenny is first, Ellen is second, Michael is third, and so on, race after race, then differences in performance are largely due to individual differences. In this case, the Between-subjects Mean Square will be large, as will the ICC, whereas the Mean Square for Error will be small.

Can the intraclass correlation be negative?

In theory, no. Values for ICC range between zero and one. The familiar Pearson product moment correlation, Chapter 16, takes any value between -1 and $+1$. However, in practice, negative values for ICC will result if $MS_B < MS_W$.

In other words, if the within-group variability is greater than the among-group variability, then a negative ICC is possible. Small ICC values and few repeats increases the risk of negative ICC estimates. Thus, a negative ICC would be “simply a(n) “unfortunate” estimate (Liljequist et al 2019).

ICC Example

I extracted 15 data points from a figure about nitrogen metabolism in kidney patients following treatment with antibiotics (Figure 1, Mitch et al. 1977). I used a web application called WebPlot Digitizer (https://apps.automeris.io/wpd/), but you can also accomplish this task within R via the `digitize` package. I was concerned about how steady my hand was using my laptop’s small touch screen, a problem that very much can be answered by thinking statistically, and taking advantage of the ICC. So, rather than taking just one estimate of each point, I repeated the protocol for extracting the points from the figure three times, generating a total of three points for each of the 15 data points (45 points in all). How consistent was I?

Let’s look at the results just for three points, #1, 2, and 3.

In the R script window enter

```
points = c(1,2,3,1,2,3,1,2,3)
```

Change `points` to character so that the ANOVA command will treat the numbers as factor levels.

```
points = as.character(points)
extracts = c(2.0478, 12.2555, 16.0489, 2.0478, 11.9637, 16.0489, 2.0478, 12.2555, 16.
```

Make a data frame, assign to an object, e.g., "digitizer"

```
digitizer = data.frame(points, extracts)
```

The dataset "digitizer" should now be attached and available to you within Rcmdr. Select digitizer data set and proceed with the one-way ANOVA.

Output from oneway ANOVA command:

```
Model.1 <- aov(extracts ~ points, data=digitizer)

 summary(Model.1)
            Df    Sum Sq   Mean Sq  F value     Pr(>F)
points       2 313.38895 156.69448 16562.49 5.9395e-12 ***
Residuals    6   0.05676   0.00946
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> numSummary(digitizer$extracts , groups=digitizer$points,
+   statistics=c("mean", "sd"))
         mean           sd data:n
1  2.04780000 0.0000000000      3
2 12.15823333 0.1684708085      3
3 16.04890000 0.0000000000      3
```

End R output. We need to calculate the ICC.

$$ICC = \frac{156.69448 - \frac{0.00946}{3}}{156.69448 - \frac{0.00946}{3} + 0.00946} = 0.9999$$

I'd say that's pretty repeatable and highly precise measurement!

But is it accurate? You should be able to disentangle accuracy from precision based on our previous discussion (Chapter 3.5), but now in the context of a practical way to quantify precision.

ICC calculations in R

We could continue to calculate the ICC by hand, but better to have a function. Here's a crack at the function to calculate ICC along with a 95% confidence interval.

```
myICC <- function(m, k, dfN, dfD) {
 testMe <- anova(m)
 MSB <- testMe$"Mean Sq"[1]
 MSE <- testMe$"Mean Sq"[2]
 varB <- MSB - MSE/k
 ICC <- varB/(varB+MSE)
```

```
 fval <- qf(c(.025), df1=dfN, df2=dfD, lower.tail=TRUE)
 CI = (k*MSE*fval)/(MSB+MSE*(k-1)*fval)
 LCIR = ICC-CI
 UCIR = ICC+CI
 myList = c(ICC, LCIR, UCIR)
 return(myList)
}
```

The user supplies the ANOVA model object (e.g., Model.1 from our example), k, which is the number of repeats per unit, and degrees of freedom for the among groups comparison (2 in this example), and the error mean square (6 in this case). Our example, run the function

```
m2ICC = myICC(Model.1, 3, 2,6); m2ICC
```

and R returns

```
[1] 0.9999396 0.9999350 0.9999442
```

with the ICC reported first, 0.9999396, followed by the lower limit (0.9999350) and the upper limit (0.9999442) of the 95% confidence interval.

In lieu of your own function, at several packages available for R will calculate the intraclass correlation coefficient and its variants. These packages are: `irr` , `psy` , `psych` , and `rptR` . For complex experiments involving multiple predictor variables, these packages are helpful for obtaining the correct ICC calculation (cf Shrout and Fleiss 1979; McGraw and Wong 1996). For the one-way ANOVA it is easier to just extract the information you need from the ANOVA table and run the calculation directly. We do so for a couple of examples.

Example: Are marathon runners consistent more consistent than my commute times?

A marathon is 26 miles, 385 yards long (42.195 kilometers). And yet, tens of thousands of people choose to run in these events. For many, running a marathon is a one-off, the culmination of a personal fitness goal. For others, it's a passion and a few are simply extraordinary, elite runners who can complete the courses in 2 to 3 hours (Table 12.3.3). That's about 12.5 miles per hour. For comparison, my 20-mile commute on the H1 freeway on Oahu typically takes about 40 minutes to complete, or 27 miles per hour (Table 12.3.2; yes, I keep track of my commute times, per Galton's famous maxim: "Whenever you can, count").

Table 12.3.2. A sampling of commute speeds, miles per hour (mph), on the H1 freeway during Dr. D's morning commute

Monday	Tuesday	Wednesday	Thursday	Friday
28.5	23.8	28.5	30.2	26.9
25.8	22.4	29.3	26.2	27.7
26.2	22.6	24.9	24.2	34.3
23.3	26.9	31.3	26.2	30.2

Calculate the ICC for my commute speeds.

Run the one-way ANOVA to get the necessary mean squares and input the values into our ICC function. We have

```
require(psych)
m2ICC = myICC(AnovaModel.1, 4, 4,11); m2ICC
[1] 0.7390535 0.6061784 0.8719286
```

Repeatability, as estimated by the ICC, was 0.74 (95% CI 0.606, 0.872), for repeat measures of commute times.

We can ask the same about marathon runners — how consistent from race to race are these runners? The following data are race times drawn from a sample of runners who completed the Honolulu Marathon in both 2016 and 2017 in 2 to 3 hours (times recorded in minutes). In other words, are elite runners consistent?

Table 12.3.3. Honolulu marathon running times (in min.) for eleven repeat, elite runners.

ID	Time 1	Time 2
P1	179.9	192.0
P2	129.9	130.8
P3	128.5	129.6
P4	179.4	179.7
P5	174.3	181.7
P6	177.2	176.2
P7	169.0	173.4
P8	174.1	175.2
P9	175.1	174.2
P10	163.9	175.9
P11	179.3	179.8

After running a one-way ANOVA, here are results for the marathon runners:

```
m2ICC = myICC(Model.1, 2, 10,11); m2ICC
[1] 0.9780464 0.9660059 0.9900868
```

Repeatability, as estimated by the ICC, was 0.98 (95% CI 0.966, 0.990), for repeat measures of marathon performance. Put more simply, knowing what a runner did in 2016 I would be able to predict their 2017 race performance with high confidence, 98%!

And now, we compare: the runners are more consistent!

Clearly this is an apples-to-oranges comparison, but it gives us a chance to think about how we might make such comparisons. The ICC will change because of differences among individuals. For example, if individuals are not variable, then xx too little variation.

An example for you to work, from our Measurement Day

If you recall, we had you calculate length and width measures on shells from samples of gastropod and bivalve species. In the table are repeated measures of shell length, by caliper in mm, for a sample of *Conus* shells (Fig. 12.3.1 and Table 12.3.4).

Figure 12.3.1: Conus shells, image by M. Dohm.

Table 12.3.4. Unstacked dataset of repeated length measures on 12 shells.

Sample	Measure 1	Measure 2	Measure 3
1	45.74	46.44	46.79
2	48.79	49.41	53.36
3	52.79	53.45	53.36
4	52.74	53.14	53.14
5	53.25	53.45	53.15
6	53.25	53.64	53.65
7	31.18	31.59	31.44
8	40.73	41.03	41.11
9	43.18	43.23	43.2
10	47.10	47.64	47.64
11	49.53	50.32	50.24
12	53.96	54.50	54.56

Questions

1. Consider data in Table 12.3.2, Table 12.3.3, and Table 12.3.4 True or False: The arithmetic mean is an appropriate measure of central tendency. Explain your answer.
2. Enter the shell data into R; it's best to copy and stack the data in your spreadsheet, then import into R or R Commander. Once imported, don't forget to change Sample to character, otherwise R will treat Sample as ratio scale data type. Run your one-way ANOVA and calculate the intraclass correlation (ICC) for the dataset. Is the shell length measure repeatable?
3. True or False. A fixed effects ANOVA implies that the researcher selected levels of all treatments.
4. True or False. A random effects ANOVA implies that the researcher selected levels of all treatments.

5. A clinician wishes to compare the effectiveness of three competing brands of blood pressure medication. She takes a random sample of 60 people with high blood pressure and randomly assigns 20 of these 60 people to each of the three brands of blood pressure medication. She then measures the decrease in blood pressure that each person experiences. This is an example of (select all that apply)
   A. a completely randomized experimental design
   B. a randomized block design
   C. a two-factor factorial experiment
   D. a random effects or Type II ANOVA
   E. a mixed model or Type III ANOVA
   F. a fixed effects model or Type I ANOVA
6. A clinician wishes to compare the effectiveness of three competing brands of blood pressure medication. She takes a random sample of 60 people with high blood pressure and randomly assigns 20 of these 60 people to each of the three brands of blood pressure medication. She then measures the blood pressure before treatment and again 6 weeks after treatment for each person. This is an example of (select all that apply)
   A. a completely randomized experimental design
   B. a randomized block design
   C. a two-factor factorial experiment
   D. a random effects or Type II ANOVA
   E. a mixed model or Type III ANOVA
   F. a fixed effects model or Type I ANOVA
7. The advantage of a randomized block design over a completely randomized design is that we may compare treatments by using _____ experimental units.
   A. randomly selected
   B. the same or nearly the same
   C. independent
   D. dependent
   E. All of the above

12.4: ANOVA from "sufficient statistics"

Introduction

By now you should be able to run a one-way ANOVA using R (and R Commander) with ease. As a reminder, You should also be aware that, if you need to, you could use spreadsheet software like Microsoft Excel or LibreOffice Calc to run a one-way ANOVA on a small data set. Still, there are times when you may need to run a one-way ANOVA on a small data set, and doing so by hand calculator may be just as convenient. What are your available options?

Following the formulas I have given would be one way to calculate ANOVA by hand, but it would be tedious and subject to error. Instead of working with the standard formulas, **calculator shortcuts** can be derived with a little algebra, and this is where I want to draw your attention now. This technique will come in handy in lab classes or other scenarios where you collect some data among a set number of groups and calculate means and standard deviations. The purpose of this posting is to show you how to obtain the necessary statistics to calculate a one-way ANOVA from the available descriptive statistics: means, standard deviations, and sample sizes. In other words, these are the **sufficient statistics** for one-way ANOVA.

> **Note:**
>
> In Chapter 11.5, we introduced use of **summary statistics**, i.e., "sufficient statistics," to calculate the independent sample t-test.

As you recall, a one-way ANOVA yields a single F test of the null hypothesis that all group means are equal. To calculate the F test, you need

- Mean Square Between Groups, MS_B
- Mean Squares Within Groups or Error, MS_E

F is then calculated as

$$F = \frac{MS_B}{MS_E}$$

with degrees of freedom \(k - 1\) for the numerator and $N - 1$ for the denominator. MS_E can also appear as MS_W.

We can calculate MS_B as

$$MS_B = \frac{\sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}_G\right)^2}{k-1}$$

where k is the number of i^{th} groups, n_i is the sample size of the i^{th} group, $\bar{X}_G$ refers to the overall mean for all of the $\bar{X}_i$ sample means.

Next, for the Error Mean Square, MS_E, all we need is the average of the sample variances (the square of the sample standard deviation, s).

$$MS_E = \frac{\sum_{i=1}^{k} s_i^2}{k}$$

ANOVA from sufficient statistics

Consider an example data set (Table 12.4.1) for which only summary statistics are available (mean and standard deviation, sd). The data set is for metabolic rate (ml oxygen per hour) for strains of laboratory mice. Sample size for each group was seven mice.

Table 12.4.1. *Descriptive statistics wheel-running behavior mice from ten different inbred strains of mice (Mus domesticus).*

Strain	n	Mean	sd
AKR	7	395	169.7
C57BL_10	7	1135	63.6
CBA	7	855	77.8
129S1	7	1012	176.8
C3H/He	7	833	49.5
C57BL/6	7	1075	91.9
FVB/N	7	1023	91.9
A	7	806	134.4
BALB/c	7	936	70.7
DBA/2	7	872	49.5

Spreadsheet calculations

You have several options at this point, ranging from using your calculator and the formulas above (don't forget to square the standard deviation to get the variances!), or you could use Microsoft Excel or LibreOffice Calc and enter the necessary formulas by hand (Table 12.4.2). You'll also find many online calculators for one-way ANOVA by sufficient statistics (e.g., https://www.danielsoper.com/statcalc/calculator.aspx?id=43).

Table 12.4.2. *Spreadsheet with formulas for calculating one-way ANOVA from means and standard deviations from statistics presented in* Table 12.4.1.

	A	B	C	D	E	F	G	H	I
1	**Strain**	**n**	**Mean**	**sd**	**squared**	**variance**		grand mean	=AVERAGE(C:
2	AKR	7	395	169.7	=B2*(C2-$I$1)^2	=D2^2		dfB	=COUNT(B:B)
3	C57BL_10	7	1135	63.6	=B2*(C3-$I$1)^2	=D3^2		dfE	=SUM(B:B)-I2
4	CBA	7	855	77.8	=B2*(C4-$I$1)^2	=D4^2		Msb	=SUM(E:E)/(
5	129S1	7	1012	176.8	=B2*(C5-$I$1)^2	=D5^2		Mse	=SUM(F:F)/C
6	C3H/He	7	833	49.5	=B2*(C6-$I$1)^2	=D6^2		F	=I4/I5
7	C57BL/6	7	1075	91.9	=B2*(C7-$I$1)^2	=D7^2		P-value	=FDIST(I6,I
8	FVB/N	7	1023	91.9	=B2*(C8-$I$1)^2	=D8^2			
9	A	7	806	134.4	=B2*(C9-$I$1)^2	=D9^2			
10	BALB/c	7	936	70.7	=B2*(C10-$I$1)^2	=D10^2			
11	DBA/2	7	872	49.5	=B2*(C11-$I$1)^2	=D11^2			

For this example, you should get the following:

MS_B = 299943.5

MS_E = 11500.8

F = 26.08

P-value = 9.75E-18

Note: The number of figures reported for the P-value implies a precision that the data simply do not support. For a report, recommend writing the P-value < 0.001

But, R can do it better.

Here's how. Install the `HH` package (or `RcmdrPlugin.HH` for use in Rcmdr) and call the `aovSufficient` function.

Step 1. Install the `HH` package from a CRAN mirror, e.g., cloud.r-project.org, in the usual way.

```
chooseCRANmirror()
install.packages("HH")
library(HH)
```

Step 2. Enter the data. Do this in the usual way (e.g., from a text file), or enter directly using the `read.table` command as follows.

```
MouseData <- read.table(header=TRUE, sep = "", text=
"Strain Mean sd
AKR 395 169.7
C57BL_10 1135 63.6
CBA 855 77.8
```

```
129S1 1012 176.8
C3H/He 833 49.5
C57BL/6 1075 91.9
FVB/N 1023 91.9
A 806 134.4
BALB/c 936 70.7
DBA/2 872 49.5")
#Check import
head(MouseData)
```

End of R input

I know, a little hard to read, but from the `MouseData` to the end bracket ") before the comment line `#Check import`, that's all one command.

Of course, you could copy the data and import the data from your computer's clipboard in **Rcmdr: Data → Import data → from text file, clipboard, or URL…** (Hint: for field separator, try White space; if that fails, try Tabs).

Once the data set is loaded, proceed to Step 3.

Step 3. In our example, sample size is included for each group. Skip to step 4. If, however, the table lacked the sample size information, you can always add a new variable. For example, if we needed to add sample size to the data frame, we would use the repeat element function, `rep()`.

```
MouseData$n <- rep(7, 10)
```

If you check the View data set button in `Rcmdr`, you will see that the command in Step 3 has added a new variable "n" for each of the eleven rows. The function `rep()` stands for "replicate elements in vectors" and what it did here was enter a value of 7 for each of the ten rows in the data set. Again, this step is not necessary for this example because sample size is already part of the data frame. Proceed to step 4.

Step 4. Run the one way ANOVA using the sufficient statistics and the `HH` function `aovSufficient`

```
MouseData.aov <- aovSufficient(Mean ~ Strain, data=MouseData)
```

Step 5. Get the ANOVA table.

```
summary(MouseData.aov)
```

Here's the R output:

```
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
Strain       9 2699491   299943    26.08  <2e-16 ***
Residuals   60  690046    11501
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

End R output.

To explore other features of the package, type `?aovSufficient` at the R prompt (like all R functions, extensive help is generally available for each function in a package).

Limitations of ANOVA from sufficient statistics

This was pretty easy, so it is worth asking — Why go through the bother of analyzing the raw data, why not just go to the summary statistics and run the calculator formula? First, the chief reason against the calculator formula and use of only sufficient statistics loses information about the individual values and therefore you have no access to the residuals. The residual of an observation is the difference between the original observation and the model prediction. The residuals are important for determining whether the model fits the data well and are, therefore, part of the toolkit that statisticians need to do proper data analysis. We will spend considerable time looking at residual patterns and it is an important aspect of doing statistics correctly.

Secondly, while it is possible to extend this approach to more complicated ANOVA problems like the **two-way ANOVA** (Cohen 2002), the statistical significance of the interaction term(s) calculated in this way are only approximate (the main effects are OK to interpret). Thus, ANOVA from sufficient statistics has its place when all you have is access to descriptive statistics, but its use is limited and not at all the preferred option for data analysis when the original, raw observations are in hand.

Questions

1. Under what circumstances would you use “sufficient statistics” to calculate a one-way ANOVA?

2. Calculate the one-way ANOVA for body weight of 47 female (F) and 97 male (M) cats (kilograms, dataset `cats` in `MASS` R package) from the following summary statistics.

	n	Mean	sd
F	47	2.36	0.274
M	97	2.9	0.468

3. Bonus: Load the `cats` data set (package `MASS`, loaded with `Rcmdr`) and run a one-way ANOVA using the `aov()` function via `Rcmdr`. Are the ANOVA from sufficient statistics the same as results from the direct ANOVA calculation? If not, why not.

12.5: Effect size for ANOVA

Introduction

As noted in the t-test chapter and our discussion of statistical power, an **effect size** is a measure of the strength of a phenomenon. An effect size calculated from data is a descriptive statistic that indicates how large (or small) the difference is between two or more samples. Effect size measures help to provide context between statistical significance, i.e., p-values. Among other uses, effect size measures play an important role in meta-analysis studies biological significance (i.e., importance of the measured difference).

Estimation of effect size

eta-squared (η^2)

$$\eta^2 = \frac{SS_B}{SS_T}$$

where SS_B and SS_T are from the one-way ANOVA and refer to the **sum of squared among (between) groups** and the **total sum of squares**, respectively. η^2 measures the proportion of the variation in the response variable that can be explained by membership in one of the groups.

For example, η^2 of 0.15 or 15% is interpreted to mean that just 15% of the variation in the response variable can be attributed to membership in the groups (i.e., whether a subject was in control group or treatment group, only 15% of the differences between individuals can be attributed to having received control).

> **Note:**
>
> η^2 is just the unadjusted coefficient of determination, R^2.

omega-squared (ω^2)

A similar, but "less biased" effect size estimate is given by omega-squared (ω^2) (Okada 2013). (The bias comes in because sample estimates were used.) Interpretation of ω^2 is the same as η^2: measures the proportion of variation explained by membership in the groups. ω^2 is given as

$$\omega^2 = \frac{SS_B - (k-1)MS_W}{SS_T + MS_W}$$

The bias for η^2 is more pronounced with small sample size, so omega-squared is preferred. ω^2 will always be less than or equal to η^2 (Okada 2013).

R code

If you have not already done so, please install the package `effectsize`.

We'll use the example one-way ANOVA problem from Chapter 12.2. I've added the data as `example.ch12` and some R script to load the data; scroll to bottom of this page or click on the link).

Get η^2

First, get the ANOVA model. There are several ways to do this within R (and Rcmdr), the most general is to use the general linear model function, `lm()`,

```
#Get the model
AnovaModel.4 <- lm(Difference~Group, data=example.ch12)
#turn it into format that the next command needs
aov_fit <- anova(AnovaModel.4)
#get eta
effectsize(aov_fit)
```

R output:

```
Parameter | Eta2 (partial) | 9e+01% CI
----------------------------------------
Group | 0.95 | [0.88, 0.97]
```

η^2 is 0.95 for the Difference group.

Get ω^2

R code:

```
omega_squared(aov_fit, partial = FALSE, ci = 0.95)
```

R output:

```
Parameter | Omega2 | 1e+02% CI
---------------------------------
Group | 0.93 | [0.82, 0.97]
```

This is a case of a large effect due to group membership. The differences between A, B, and C means account for 93% of the variation. Note that both η^2 and ω^2 are close; note also that η^2 is greater than ω^2.

Questions

1. In Chapter 7.4 we introduced the concept of number needed to treat, NNT. Discuss the concept of "important differences" between sample means of a control group and a treatment group to NNT.
2. For the "No difference" group, calculate η^2 and ω^2. Use the box plots shown in Figure $12.2.1$ together with effect size statistics to discuss the relationship between statistical significance (p-value) and important difference.

Data set

```
example.ch12 <- read.table(header=TRUE, sep=",",text="
Group, No.difference, Difference
A, 12.04822, 11.336161
A, 12.67584, 13.476142
A, 12.99568, 12.961210
A, 12.01745, 11.746712
A, 12.34854, 11.275492
B, 12.17643, 7.167262
B, 12.77201, 5.136788
B, 12.07137, 6.820242
B, 12.94258, 5.318743
B, 12.07670, 7.153992
C, 12.58212, 3.344218
C, 12.69263, 3.792337
C, 12.60226, 2.444438
C, 12.02534, 2.576014
C, 12.60420, 4.575672")
#check the dataframe
head(example.ch12)
```

12.6: ANOVA post-hoc tests

ANOVA post-hoc tests

Tests of the null hypothesis in a one-way ANOVA yields one answer: either you reject the null, or you do not reject the null hypothesis.

But while there was only one factor (population, drug treatment, etc) in a one-way ANOVA, there are usually many treatments (e.g., multiple levels, four different populations, 3 doses of a drug plus a placebo). ANOVA plus **post-hoc tests** solves the multiple comparison problem we discussed: you still get your tests of all group differences, but with adjustments to the procedures so that these tests are conducted without suffering the increase in type I error = the multiple comparison problem. If the null hypothesis is rejected, you may then proceed to post-hoc tests among the groups to identify differences.

Consider the following example of four populations scored for some outcome, `sim.ch12` (scroll down the page, or click here to get the R code).

Bring the data frame, `sim.ch12` , into current memory in Rcmdr by selecting the data set. Next, run the one-way ANOVA.

Rcmdr: Statistics → Means → One-way ANOVA…

which brings up the following menu (Fig. 12.6.1)

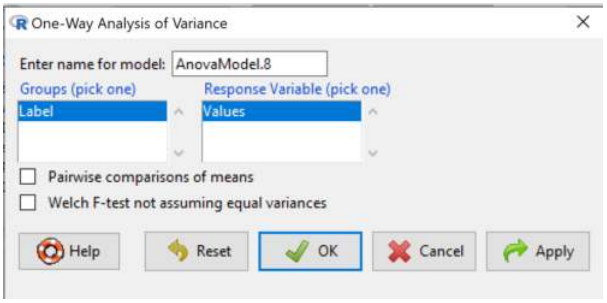


Figure 12.6.1: One-way ANOVA menu in R Commander.

> **Note:**
>
> If you look carefully in Figure 12.6.1, you can see model name was `AnovaModel.8` . There's nothing significant about that name, it just means this was the 8th model I had run up to that point. As a reminder, Rcmdr will provide names for models for you; it is better practice to provide model names yourself.

Notice that Rcmdr menu correctly identifies the Factor variable, which contains text labels for each group, and the Response variable, which contains the numerical observations.

> **Note:**
>
> If your factor is numeric, you'll first have to tell R that the variable is a factor and hence nominal. this can be accomplished within Rcmdr via the Data Manage variables… options, or simply submit the command
>
> ```
> newName <- as.factor(oldVariable)
> ```

If your data set contains more variables, then you would need to sort through these and select the correct model (Fig. 12.6.1).

To get the default Tukey post-hoc tests simply check the Pairwise comparisons box and then click OK.

For a test of the null that four groups have the same mean, a publishable **ANOVA table** would look like…

Table 12.6.1. *The ANOVA table.*

	Df	Mean Square	F	P†
Label	3	389627	76.44	< 0.0001

	Df	Mean Square	F	P†
Error	36	61167		

† Dr. D edited the R output for p-value. R doesn't report P as less than some value.

Note:

The ANOVA table is something you put together from the output of R (or other statistical programs).

Here's the R output for ANOVA:

```
AnovaModel.8 <- aov(Values ~ Label, data = sim.Ch12)
summary(AnovaModel.8)
            Df  Sum Sq Mean Sq  F value     Pr(>F)
Label        3  389627  129876    76.44   1.11e-15 ***
Residuals 36   61167    1699
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
with(sim.Ch12, numSummary(Values, groups = Label, statistics = c("mean", "sd")))
      mean       sd data:n
Pop1 150.3 35.66838 10
Pop2  99.0 43.25634 10
Pop3 130.9 42.36469 10
Pop4 350.7 43.10723 10
```

End of R output.

Recall that all we can say is that a difference has been found and we reject the null hypothesis. However, we do not know if group 1 = group 2, but both are different from group 3, or some other combination. So we need additional tools. We can conduct post-hoc tests (also called multiple comparisons tests).

Once a difference has been detected (F test statistic > F critical value, therefore $P < 0.05$), then ***posteriori tests***, also called **unplanned comparisons**, can be used to tell which means differ.

There are also cases for which some comparisons were planned ahead of time and these are called ***a priori*** or **planned comparisons**; even though you conduct the tests after the ANOVA, you were always interested in particular comparisons. This is an important distinction: planned comparisons are more powerful, more aligned with what we understand to be the scientific method.

Let's take a look at these procedures. Collectively, they are often referred to as **post-hoc** tests (Ruxton and Beauchamp 2008). There are many different flavors of these tests, and R offers several, but I will hold you responsible only for three such comparisons: **Tukey's, Dunnett's, and Bonferroni (Dunn)**. These named tests are among the common ones, but you should be aware that the problem of multiple comparisons and inflated error rates has received quite a lot of recent attention because the size of data sets has increased in many fields, e.g., genome wide-association studies in genetics or data mining in economics or business analytics. A related topic then is the issue of "false positives." New approaches include Holm-Bonferroni. There are others — it is a regular "cottage industry" in applied statistics to a problem that, while recognized, has not achieved a universal agreed solution. Best we can do is be aware and deal with it and know that the problem is one mostly of big data (e.g., microarray and other high-through put approaches).

Important R Note: In order to do most of the post-hoc tests you will need to install the `multcomp` package; after installing the package, load the `library(multcomp)` . Just using the default option from the one-way ANOVA command yields the Tukey's HSD test.

Performing multiple comparisons and the one-way ANOVA

a. Tukey's: "honestly (wholly) significant difference test"

Tests $H_O : \bar{X}_B = \bar{X}_A$ versus $H_A : \bar{X}_B \neq \bar{X}_A$ where A and B can be any pairwise combination of two means you wish to compare. There are $\frac{k(k-1)}{2}$ comparisons.

$$q = \frac{\bar{X}_B - \bar{X}_A}{SE}$$

where

$$SE = \sqrt{\frac{MS_{error}}{n}}$$

and n is the harmonic mean of the sample sizes of the two groups being compared. If the sample sizes are equal, then the simple arithmetic mean is the same as the harmonic mean.

Note:

q is like t for when we are testing means from two samples.

- The significance level is the probability of encountering at least one Type I error (probability of rejecting H_O when it is true). This is called the **experiment-wise (family-wise) error rate** whereas before we talked about the **comparison-wise (individual) error rate**.

Two options to get the post-hoc test Tukey — use a package called mcp or in Rcmdr, Tukey is the default option in the one-way ANOVA command.

Rcmdr: Statistics → Means → One-way ANOVA
Check "Pairwise comparisons of means" to get the Tukey'a HSD test (Fig. 12.6.2)

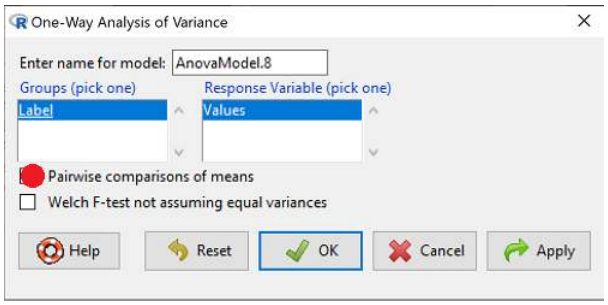


Figure 12.6.2: Select Tukey post-hoc tests with the one-way ANOVA.

R output follows. There's a lot, but much of it is repeat information. Take your time, here we go.

```
.Pairs <- glht(AnovaModel.4, linfct = mcp(Label = "Tukey"))
summary(.Pairs) # pairwise tests
     Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
 Fit: aov(formula = Values ~ Label, data = sim.Ch12)

 Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
 Pop2 - Pop1 == 0   -51.30      18.43  -2.783   0.0405 *
 Pop3 - Pop1 == 0   -19.40      18.43  -1.052   0.7201
 Pop4 - Pop1 == 0   200.40      18.43  10.871   <0.001 ***
 Pop3 - Pop2 == 0    31.90      18.43   1.730   0.3233
```

```
Pop4 - Pop2 == 0   251.70      18.43  13.654   <0.001 ***
Pop4 - Pop3 == 0   219.80      18.43  11.924   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

confint(.Pairs) # confidence intervals
     Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Values ~ Label, data = sim.Ch12)

Quantile = 2.6927
95% family-wise confidence level

Linear Hypotheses:
                  Estimate  lwr       upr
Pop2 - Pop1 == 0  -51.3000 -100.9382   -1.6618
Pop3 - Pop1 == 0  -19.4000  -69.0382   30.2382
Pop4 - Pop1 == 0  200.4000  150.7618  250.0382
Pop3 - Pop2 == 0   31.9000  -17.7382   81.5382
Pop4 - Pop2 == 0  251.7000  202.0618  301.3382
Pop4 - Pop3 == 0  219.8000  170.1618  269.4382
```

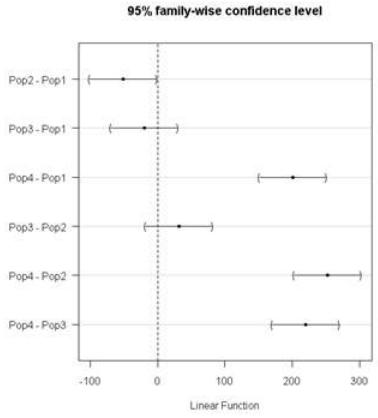


Figure 12.6.3: Plot of confidence intervals of Tukey HSD.

R Commander includes a default 95% CI plot (Fig. 12.6.3). From this graph, you can quickly identify the pairwise comparisons for which 0 (zero, dotted vertical line) is included in the interval, i.e., there is no difference between the means (e.g., Pop1 is different from Pop4, but Pop1 is not different from Pop3).

b. Dunnett's Test for comparisons against a control group

- There are situations where we might want to compare our experimental Populations to one control Population or group.
- This is common in medical research where there is a placebo (control pill with no drug) or sham operations (operations where every thing but the critical operation is done).
- This is also a common research design in ecological or agricultural research where some animal or plant populations are exposed to an environmental factor (e.g. fertilizer, pesticide, pollutant, competitors, herbivores) and other animal or plant populations are not exposed to these environmental factor.

- The difference in the statistical procedure for analyzing this type of research design is that the experimental groups may only be compared to the control group.
- This results in fewer comparisons.
- The formula is the same as for the Tukey's Multiple Comparison test, except for the calculation of the SE.

Standard Error is changed by multiplying the MS_{Error} by 2.

And n is the harmonic mean of the sample sizes of the two groups being compared.

$$q = \frac{\bar{X}_{control} - \bar{X}_A}{SE}$$

where

$$SE = \sqrt{\frac{2/MS_{error}}{n}}$$

R Commander doesn't provide a simple way to get Dunnett, but we can get it simply enough if we are willing to write some script. Fortunately (OK, by design!), Rcmdr prints commands.

Look at the Output window from the one-way ANOVA with pairwise comparisons: it provides clues as to how we can modify the `mcp` command (`mcp` stands for multiple comparisons).

First, I had run the one-way ANOVA command and noted the model (`AnovaModel.3`). Second, I wrote the following script, modified from above.

```
Pairs <- glht(AnovaModel.1, linfct = mcp(Label = c("Pop2 - Pop1 = 0", "Pop3 - Pop1 =
```

where Label is my name for the Factor variable. Note that I specified the comparisons I wanted R to make. When I submit the script, nothing shows up in the Output window because the results are stored in my "Pairs."

I then need to ask R to provide confidence intervals

```
confint(Pairs)
```

R output window

```
Pairs <- glht(AnovaModel.1, linfct = mcp(Label = c("Pop2 - Pop1 = 0", "Pop3 - Pop1 =
confint(Pairs)

Simultaneous Confidence Intervals

 Multiple Comparisons of Means: User-defined Contrasts

 Fit: aov(formula = Values ~ Label, data = sim.Ch12)

 Estimated Quantile = 2.4524
 95% family-wise confidence level

 Linear Hypotheses:
  ...... lwr ....... upr
 Pop2 - Pop1 == 0  .. -51.3000  . -96.5080  ... -6.0920
 Pop3 - Pop1 == 0  .. -19.4000  . -64.6080  ... 25.8080
 Pop4 - Pop1 == 0  .. 200.4000  . 155.1920  .. 245.6080
```

Look for intervals that include zero, therefore, the group does not differ from the Control group (Pop1). How many groups differed from the Control group?

Alternatively, I may write

```
Tryme <- glht(AnovaModel.1, linfct = mcp(Label = "Dunnett"))
confint(Tryme)
```

It's the same (in fact, the default mcp test is the Dunnett).

```
Tryme <- glht(AnovaModel.1, linfct = mcp(Label = "Dunnett"))
confint(Tryme)

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = Values ~ Label, data = sim.Ch12)

Estimated Quantile = 2.4514
95% family-wise confidence level

Linear Hypotheses:
Estimate lwr upr
Pop2 - Pop1 == 0 -51.3000 -96.4895 -6.1105
Pop3 - Pop1 == 0 -19.4000 -64.5895 25.7895
Pop4 - Pop1 == 0 200.4000 155.2105 245.5895
```

c. Bonferroni t

The *Bonferroni t* test is a popular tool for conducting multiple comparisons. The rationale for this particular test is that the MS_{error} is a good estimate of the pooled variances for all groups in the ANOVA.

$$Bonferroni = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{MS_{error}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}}$$

and $DF = N - k$.

Note:

In order to achieve a Type I error rate of 5% for all tests, you must divide the 0.05 by the number of comparisons conducted.

Thus, for $k = 4$ groups, $\binom{4}{2} = \frac{4!}{2!(4-2)!}$

Here's a more general version if you prefer to get all pairwise tests: $\binom{k}{2} = \frac{k!}{2!(k-2)!}$

Use this information then to determine how many total comparisons will be made, then if necessary, use to adjust Type I error rate for one test (the exeriment-wise error rate).

For our example, the adjusted Type I error is $0.05/6 = 0.00833$ Thus, for a difference between two means to be statistically significant, the P-value must be less than 0.00833.

For Bonferroni, we will use the following script.

1. Set up one-way ANOVA model (ours has been saved as AnovaModel.1),
2. Collect all pairwise comparisons with the mcp(~"Tukey") stored in a vector (I called mine Whynot),
3. and finally, get the Bonferroni adjusted test of the comparisons with the summary command, but add the "test = adjusted("bonferroni").

It's a bit much, but we end up with a very nice output to work with.

```
Whynot <- glht(AnovaModel.3, linfct = mcp(Label = "Tukey"))
summary(Whynot, test = adjusted("bonferroni"))

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Values ~ Label, data = sim.Ch12)

Linear Hypotheses:
Estimate Std. Error t value Pr(>|t|)
Pop2 - Pop1 == 0 -51.30 18.43 -2.783 0.0512 .
Pop3 - Pop1 == 0 -19.40 18.43 -1.052 1.0000
Pop4 - Pop1 == 0 200.40 18.43 10.871 3.82e-12 ***
Pop3 - Pop2 == 0 31.90 18.43 1.730 0.5527
Pop4 - Pop2 == 0 251.70 18.43 13.654 5.33e-15 ***
Pop4 - Pop3 == 0 219.80 18.43 11.924 2.77e-13 ***
  ---
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 (Adjusted p values reported -- bonferroni method)
```

Questions

1. Be able to define and contrast experiment-wise and family-wise error rates.
2. Read and interpret R output
    1. Refer back to the Tukey HSD output. Among the four populations, which pairwise groups were considered "statistically significant" following use of the Tukey HSD?
    2. Refer back to the Dunnett's output. Among the four populations, which population was taken as the control group for comparison?
    3. Refer back to the Dunnett's output. Which pairwise groups were considered "statistically significant" from the control group?
    4. Refer back to the Bonferroni output. Among the four populations, which pairwise groups were considered "statistically significant" following use of the Bonferroni correction?
    5. Compare and contrast interpretation of results for post-hoc comparisons among the four populations based on the three different post-hoc methods
3. Be able to distinguish when Tukey HSD and Dunnet's post hoc tests are appropriate.
4. Some microarray researchers object to use of Bonferroni correction because it is too "conservative." In the context of statistical testing, what errors are the researchers talking about when they say the correction is "conservative"?

Data set used in this page

```
sim.ch12 <- read.table(header=TRUE, sep=",",text="
Label, Value
```

```
Pop1, 105
Pop1, 132
Pop1, 156
Pop1, 198
Pop1, 120
Pop1, 196
Pop1, 175
Pop1, 180
Pop1, 136
Pop1, 105
Pop2, 100
Pop2, 65
Pop2, 60
Pop2, 125
Pop2, 80
Pop2, 140
Pop2, 50
Pop2, 180
Pop2, 60
Pop2, 130
Pop3, 130
Pop3, 95
Pop3, 100
Pop3, 124
Pop3, 120
Pop3, 180
Pop3, 80
Pop3, 210
Pop3, 100
Pop3, 170
Pop4, 310
Pop4, 302
Pop4, 406
Pop4, 325
Pop4, 298
Pop4, 412
Pop4, 385
Pop4, 329
Pop4, 375
Pop4, 365")
#check the dataframe
head(sim.ch12)
```

12.7: Many tests, one model

Introduction

In our introduction to parametric tests we so far have covered one- and two-sample t-tests and now the multiple sample or one-way analysis of variance (ANOVA). In subsequent sections we will cover additional tests, each with their own name. It is time to let you in on a little secret. All of these tests, t-tests, ANOVA, and linear and multiple regression that we will work on later in the book, belong to one family of statistical models. That model is called the general Linear Model (LM), not to be confused with the Generalized Linear Model (GLM) (Burton et al 1998; Guisan et al 2002). This greatly simplifies our approach to learning how to implement statistical tests in R (or other statistical programs) — you only need to learn one approach: the general Linear Model (LM) function `lm()` .

Brief overview of linear models

With the inventions of correlation, linear regression, t-tests, and analysis of variance in the period between 1890 and 1920, subsequent work led to the realization that these tests (and many others!) were special cases of a general model, the general linear model, or LM. The LM itself is a special case of the generalized linear model, or GLM; among the differences between LM and GLM, in LM, the dependent variable is ratio scale and errors in the **response (dependent) variable(s)** are assumed to come from a Gaussian (normal) distribution. In contrast, for GLM, the response variable may be categorical or continuous, and error distributions other than normal (Gaussian), may be applied. The GLM user must specify both the **error distribution family** (e.g., Gaussian) and the **link function**, which specifies the relationship among the response and predictor variables. While we will use the GLM functions when we attempt to model growth functions and calculate EC_{50} in dose-response problems, we will not cover GLM this semester.

The general Linear Model, LM

In matrix form, the LM can be written as $\hat{\mathbf{Y}} = \mathbf{Xb}$

where $\hat{\mathbf{Y}}$ is a matrix of response variables predicted by **independent variables** contained in matrix $\mathbf{X}$ and weighted by **linear coefficients** in the vector $\mathbf{b}$. Basically, all of the predictor variables are combined to produce a single linear predictor $\mathbf{Xb}$. By adding an error component we have the complete linear model: $\mathbf{Y} = \mathbf{X}\beta$

In the linear model, the error distribution is assumed to be normally distributed, or “Gaussian.”

R code

The bad news is that LM in R (and in any statistical package, actually) is a fairly involved set of commands; the good news is that once you understand how to use this command, and can work with the Options, you will be able to conduct virtually all of the tests we will use this semester, from two-sample t-tests to multiple linear regression. In the end, all you need is the one Rcmdr command to perform all of these tests.

We begin with a data set, `ohia.ch12` . Scroll down this page or click here to get the R code.

I found a nice report on a common garden experiment with o`hia (Corn and Hiesey 1973). O`hia (*Metrosideros polymorpha*) is an endemic, but wide-spread tree in the Hawaiian islands (Fig. 12.7.1). O`hia exhibits pronounced intraspecific variation: individuals differ from each other. O`hia grows over wide range of environments, from low elevations along the ocean right up the sides of the volcanoes, and takes on many different growth forms, from shrubs to trees. Substantial areas of o`hia trees on the Big Island are dying, attributed to two exotic fungal species of the genus *Ceratocystis* (Asner et al., 2018).

Figure 12.7.1: O’hia, *Metrosideros polymorpha*. Public domain image from Wikipedia.

The Biology. Individuals from distinct populations may differ because the populations differ genetically, or because the environments differ, or, and this is more realistic, both. Phenotypic plasticity is the ability of one genotype to produce more than one phenotype when exposed to different environments. Environmental differences are inevitable when populations are from different geographic areas. Thus, in population comparisons, genetic and environmental differences are confounded. A common garden experiment is a crucial genetic experiment to separate variation in phenotypes, P, among populations into causal genetic or environmental components.

If you recall from your genetics course, $P = G + E$ where G stands for genetic (alleles) differences among individuals and E stands for environmental differences among individuals. In brief, the common garden experiment begins with individuals from the different populations are brought to the same location to control environmental differences. If the individuals sampled from the populations continue to differ despite the common environment, then the original differences between the populations must have a genetic basis, although the actual genetic scenario may be complicated (the short answer is that if genotype by environment interaction exists, then results from a **common garden experiment** cannot be generalized back to the natural populations/locations — this will make more sense when we talk about two-way ANOVA). For more about common garden experiments, see de Villemereuil et al (2016). Nuismer and Gandon (2008) discuss statistical aspects of the common garden approach to studying local adaptation of populations and the more powerful "reciprocal translocation" experimental design.

Managing data for linear models

First, your data must be stacked in the worksheet. That means one column is for group labels (independent variable), the other column is for the response (dependent) variable.

If you have not already downloaded the data set, `ohia.ch12` , do so now. Scroll down this page or click here to get the R code.

Confirm that the worksheet is stacked. If it is not, then you would rearrange your data set using **Rcmdr: Data → Active data set → Stack variables in data set…**

The data set contains one factor, "Site" with three levels (M-1, 2, 3). M stands for Maui, and collection sites were noted in Figure 2 of Corn and Hiesey (1973). Once the dataset is in `Rcmdr` , click on View to see the data (Fig. 12.7.2). There are two response variables, Height (shown in red below) and Width (shown in blue below).

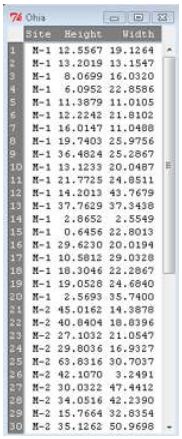


Figure 12.7.2: The o`hia dataset as viewed in R Commander.

The data are from Table 5 of Corn and Hiesey (1973). (I simulated data based on their mean/SD reported in Table 5). This was a very cool experiment: they collected o`hia seeds from three elevations on Maui, then grew the seeds in a common garden in Honolulu. Thus, the researchers controlled the environment; what varied, then were the genotypes.

As always, you should look at the data. Box plots are good to compare central tendency (Fig. 12.7.3).

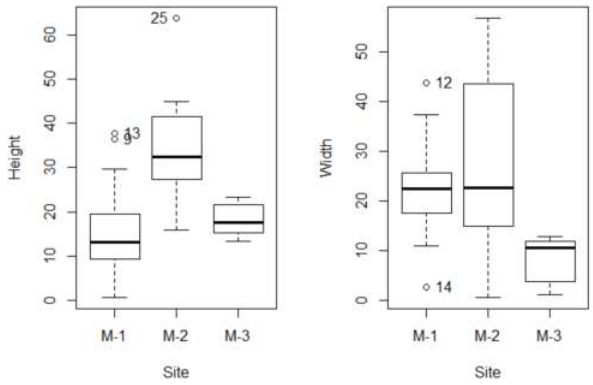


Figure 12.7.3: Box plots of growth responses of o`hia seedlings collected from three Maui sites, M-1 (elevation 750 ft), M-2 (elevation 1100 ft), and M-3 (elevation 6600 ft). Data adapted from Table 5 of Corn and Hiersey 1973.

R code to make Figure 12.7.3 plots:

```
par(mfrow=c(1,2))
Boxplot(Height ~ Site, data = ohia, id = list(method = "y"))
Boxplot(Width ~ Site, data = ohia, id = list(method = "y"))
```

This dataset would typically be described as a one-way ANOVA problem. There was one treatment variable (population source) with three levels (M-1, M-2, M-3). From `Rcmdr` we select the one-way ANOVA: **Statistics → Means → One-way ANOVA…** and after selecting the Groups (from the Site variable) and the Response variable (e.g., Height), we have

```
AnovaModel.1 <- aov(Height ~ Site, data = ohia.ch12)
summary(AnovaModel.1)
            Df Sum Sq   Mean Sq  F value          Pr(>F)
Site         2   4070    2034.8    22.63 0.000000131 ***
Residuals   47   4227      89.9
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let us proceed to test the null hypothesis (what was it???) using instead the `lm()` function. Four steps in all.

Step 1. Rcmdr: **Statistics → Fit models → Linear model …** (Fig. 12.7.4)

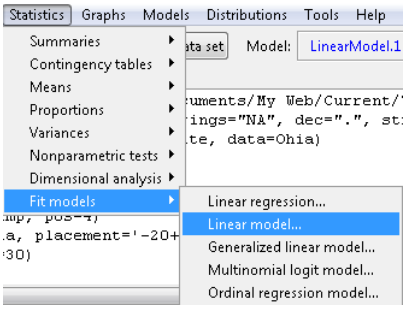


Figure 12.7.4: R Commander, select to fit a Linear model.

Step 2. The popup menu below (Fig. 12.7.5) follows.

First, What is our response (dependent) variable? What is our predictor (independent) variable? We then input our model. In this case, with only the one predictor variable, Sites, our **model formula** is simple to enter (Fig. 12.7.5): `Height ~ Site`

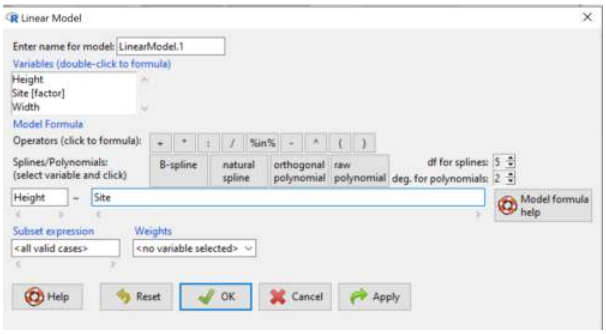


Figure 12.7.5: Input linear model formula, `Height ~ Site`

Step 3. Click OK to carry out the command.

Here is the R output and the statistical results from the application of the linear model.

```
LinearModel.1 <- lm(Height ~ Site, data=ohia.ch12)
summary(LinearModel.1)

Call:
lm(formula = Height ~ Site, data = ohia.ch12)

Residuals:
    Min      1Q  Median      3Q     Max
-18.808 -4.761  -1.755   4.758  29.257

Coefficients:
             Estimate  Std. Error  t value      Pr(>|t|)
(Intercept)    15.314       2.121    7.222 0.00000000377 ***
Site[T.M-2]    19.261       2.999    6.423 0.00000006153 ***
Site[T.M-3]     2.924       3.673    0.796          0.43
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9.483 on 47 degrees of freedom
Multiple R-squared:  0.4905,    Adjusted R-squared:  0.4688
F-statistic: 22.63 on 2 and 47 DF,  p-value: 0.0000001311
```

End R output

The linear model has produced a series of estimates of coefficients for the linear model, statistical tests of the significance of each component of the model, and the **coefficient of determination, R^2**, which is a descriptive statistic of how well model fits the data. Instead of our single factor variable for Source Population like in ANOVA we have a series of what are called **dummy variables** or **contrasts** between the populations. Thus, there is a coefficient for the difference between M-1 and M-2. " `Site[T.M-2]` " in the output, between M-1 and M-3, and between M-2 and M-3.

Note:

This is a brief description of linear model output; these topics will be discussed more fully in Chapter 17 and Chapter 18. The residual standard error is a measure of how well a model fits the data. The Adjusted R-squared is calculated by dividing the residual mean square error by the total mean square error. The result is then subtracted from 1.

It also produced our first statistic that assesses how well the model fits the data called the coefficient of determination, R^2. A R^2 value of of 1.0 would indicate that all variation in the data set can be explained by the predictor variable(s) in the model with no

residual error remaining. Our value of 49% indicates that nearly 50% of the variation in height of the seedlings grown under common environments are due to the source population (= genetics).

Step 4. But we are not quite there — we want the traditional ANOVA results (recall the **ANOVA table**).

To get the ANOVA Table we have to ask `Rcmdr` (and therefore R) to give us this. Select

Rcmdr: Models → Hypothesis tests → ANOVA table … (Fig. 12.7.6)

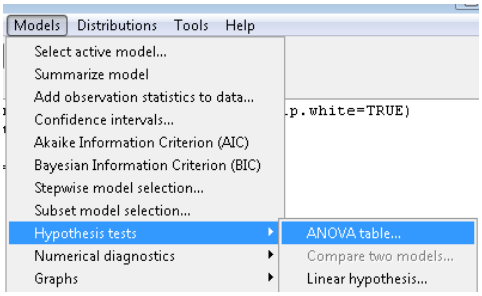


Figure 12.7.6: To retrieve an ANOVA table, select Models, Hypothesis tests, then ANOVA table…

Here's the type of tests for the ANOVA table; select the default (Fig. 12.7.7).

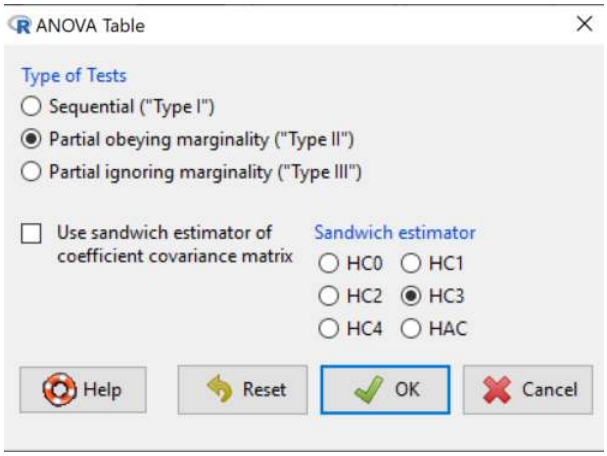


Figure 12.7.7: Options for types of tests for ANOVA table.

Now, in the future when we work with more complicated experimental designs, we will also need to tell R how to conduct the test. For now, we will accept the default **Type II type of test** and ignore **sandwich estimators.** You should confirm that for a one-way ANOVA, Type I and Type II choices give you the same results.

The reason they do is because there is only one factor — when there are more than one factors, and if one or both of the factors are random effects, then Type I, II, and III will give you different answers. We will discuss this more as needed, but see the note below about default choices.

Note:

Marginal or **partial effects** are slopes (or first derivatives): they quantify the change in one variable given change in one or more independent variables. Type I tests are sequential: sums of squares are calculated in the order the predictor variables are entered into the model. Type II tests the sums of squares as calculated after adjusting for some of the variables in the model. For Type III, every sum of square calculation is adjusted for all other variables in the model. Sandwich estimator refers to algorithms for calculating the structure of errors or residuals remaining after the predictor variables are fitted to the data. The assumption for ordinary least-square estimation (see Chapter 17) is that errors across the predictors are equal, i.e., equal variances assumption. HC refers to "heteroscedasticity consistent" (Hayes and Chai 2007).

By default, `Rcmdr` makes Type II. In most of the situations we will find ourselves this semester, this is the correct choice.

Below is the output from the ANOVA table request. Confirm that the information is identical to the output from the call to `aov()` function.

```
Anova(LinearModel.1, type = "II")
Anova Table (Type II tests)
Response: Height
          Sum Sq  Df   F value        Pr(>F)
Site      4069.7   2    22.626  0.0000001311 ***
Residuals 4226.9  47
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And the other stuff we got from the linear model command? Ignore for now but make note that this is a hint that regression and ANOVA are special cases of the same model, the linear model.

We do have some more work to do with ANOVA, but this is a good start.

Why use the linear model approach?

Chief among the reasons to use the `lm()` approach is to emphasize that a model approach is in use. One purpose of developing a model is to provide a formula to predict new values. Prediction from linear models is more fully developed in Chapter 17 and Chapter 18, but for now, we introduce the `predict()` function with our O`hia example.

```
myModel <- predict(LinearModel.1, interval = "confidence")
head(myModel, 3)    #print out first 3 rows
#Add the output to the data set
ohiaPred <- data.frame(ohia,myModel)
with(ohiaPred, tapply(fit, list(Site), mean, na.rm = TRUE))   #print out predicted va
```

Output from R

```
myModel <- predict(LinearModel.1, interval = "confidence")
head(myModel, 3)
       fit      lwr      upr
1 15.31374 11.04775 19.57974
2 15.31374 11.04775 19.57974
3 15.31374 11.04775 19.57974
with(myModel, tapply(fit, list(Site), mean, na.rm = TRUE))
     M-1      M-2      M-3
15.31374 34.57474 18.23796
```

Questions

1. Revisit ANOVA problems in homework and questions from early parts of this chapter and apply `lm()` followed by Hypothesis testing (**Rcmdr: Models → Hypothesis tests → ANOVA table**) approach instead of one-way ANOVA command. Compare results using `lm()` to results from One-way ANOVA and other ANOVA problems.

Data set and R code used in this page

Corn and Hiesey (1973). Ohia common garden.

Site	Height	Width
M-1	12.5567	19.1264
M-1	13.2019	13.1547

Site	Height	Width
M-1	8.0699	16.032
M-1	6.0952	22.8586
M-1	11.3879	11.0105
M-1	12.2242	21.8102
M-1	16.0147	11.0488
M-1	19.7403	25.9756
M-1	36.4824	25.2867
M-1	13.1233	20.0487
M-1	21.7725	24.8511
M-1	14.2013	43.7679
M-1	37.7629	37.3438
M-1	2.8652	2.5549
M-1	0.6456	22.8013
M-1	29.623	20.0194
M-1	10.5812	29.0328
M-1	18.3046	22.2867
M-1	19.0528	24.684
M-1	2.5693	35.74
M-2	45.0162	14.3878
M-2	40.8404	18.8396
M-2	27.1032	21.0547
M-2	29.8036	16.9327
M-2	63.8316	30.7037
M-2	42.107	3.2491
M-2	30.0322	47.4412
M-2	34.0516	42.239
M-2	15.7664	32.8354
M-2	35.1262	50.9698
M-2	43.6988	19.3897
M-2	26.7585	13.8168
M-2	36.7895	0.5817
M-2	30.9458	53.7757
M-2	26.8465	15.4137
M-2	40.3883	9.2161

Site	Height	Width
M-2	30.6555	56.8456
M-2	19.9736	44.9411
M-2	27.676	36.8543
M-2	44.084	24.3396
M-3	15.2646	11.4999
M-3	19.6745	9.7757
M-3	23.275	12.7825
M-3	16.1161	2.4065
M-3	16.8393	1.1253
M-3	23.107	3.7349
M-3	21.5322	6.9725
M-3	13.4191	12.2867
M-3	14.7273	11.4841
M-3	18.4245	11.9078

```
ohia.ch12 <- read.table(header=TRUE, sep=",",text="
Site, Height, Width
M-112.556719.1264
M-113.201913.1547
M-18.069916.032
M-16.095222.8586
M-111.387911.0105
M-112.224221.8102
M-116.014711.0488
M-119.740325.9756
M-136.482425.2867
M-113.123320.0487
M-121.772524.8511
M-114.201343.7679
M-137.762937.3438
M-12.86522.5549
M-10.645622.8013
M-129.62320.0194
M-110.581229.0328
M-118.304622.2867
M-119.052824.684
M-12.569335.74
M-245.016214.3878
M-240.840418.8396
M-227.103221.0547
M-229.803616.9327
```

```
M-263.831630.7037
M-242.1073.2491
M-230.032247.4412
M-234.051642.239
M-215.766432.8354
M-235.126250.9698
M-243.698819.3897
M-226.758513.8168
M-236.78950.5817
M-230.945853.7757
M-226.846515.4137
M-240.38839.2161
M-230.655556.8456
M-219.973644.9411
M-227.67636.8543
M-244.08424.3396
M-315.264611.4999
M-319.67459.7757
M-323.27512.7825
M-316.11612.4065
M-316.83931.1253
M-323.1073.7349
M-321.53226.9725
M-313.419112.2867
M-314.727311.4841
M-318.424511.90782")
#check the dataframe
head(ohia.ch12)
```

12.8: Chapter 12 References

Asner, G., Martin, R., Keith, L., Heller, W., Hughes, M., Vaughn, N., … & Balzotti, C. (2018). A spectral mapping signature for the Rapid Ohia Death (ROD) pathogen in Hawaiian forests. *Remote Sensing*, 10(3), 404.

Bewick, V., Cheek, L., Ball, J. (2004). Statistics review 9: One-way analysis of variance. *Critical Care* 8:130–136.

Bewick, V., Cheek, L., Ball, J. (2004). Statistics review 9: One-way analysis of variance. *Critical Care* 8:130-136.

Boake, C. R. B. (1989). Repeatability: its role in evolutionary studies of mating behavior. *Evolutionary Ecology* 3:173-182.

Burton, P., Gurrin, L., Sly, P. (1998). Extending the simple linear regression model to account for correlated responses : An introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine* 17:1261-1291.

Cohen, B. H. (2002). Calculating a factorial ANOVA from means and standard deviations. *Understanding Statistics* 1:191-203.

Corn, C. A., Hiesey, W. M. (1973). *Altitudinal ecotypes in Hawaiian Metrosideros*. No 10, US International Biology Program.

de Villemereuil, P., Gaggiotti, O. E., Mouterde, M., & Till-Bottraud, I. (2016). Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity*, 116(3), 249.

Dohm, M. R. (2002). Repeatability estimates do not always set an upper limit to heritability. *Functional Ecology* 16:273-280.

Guisan,A., Edwards, Jr., T. C., Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modeling* 157:89-100.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Hunter, J. E., Schmidt. F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment* 8:275-292.

Jafari, M., & Ansari-Pour, N. (2019). Why, when and how to adjust your P values?. *Cell Journal (Yakhteh)*, *20*(4), 604.

Jeanmougin, M., de Reynies, A., Laetitia, M., Paccard, C., Nuel, G., Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PloS One* 5:e12336.

Jinyuan, L. I. U., Wan, T. A. N. G., Guanqin, C. H. E. N., Yin, L. U., & Changyong, F. E. N. G. (2016). Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai archives of psychiatry, 28(2), 115.

Lessells, B., & Boag, P. T. (1987). Unrepeatable repeatabilities: A common mistake. *Auk* 104:116-121.

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PloS one*, *14*(7), e0219854.

Matthews, J. N., Altman, D. G., Campbell, M. J., Royston, P. (1990). Analysis of serial measurements in medical research. *BMJ* 300:230-235.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1), 30.

Mitch, W. E., & Walser, M. (1977). Effects of oral neomycin and kanamycin in chronic uremic patients: II. Nitrogen balance. *Kidney international*, 11(2), 123-127.

Moore, T. J., Conlin, P. R. , Ard, J., Svetkey, L. P. (2001). DASH (Dietary Approaches to Stop Hypertension) Diet Is Effective Treatment for Stage 1 Isolated Systolic Hypertension. *Hypertension* 38: 155-158.

Nakagawa, S., Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* 82:591-605.

Nuismer, S. L., Gandon, S. (2008). Moving beyond common-garden and transplant designs: Insight into the causes of local adaptation in species interactions. *American Naturalist* 171:658-668.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129-147. (cited in Daniel Lakens blog)

Ruxton, G. D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral ecology*, 19(3), 690-693.

Shrout, P. E., Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86:420-428.

Sokal, R. R., Rohlf, F. J. (1995). *Biometry: The principles and practices of statistics in biological research*, 3rd ed. WH Freeman.

Tobert, J. A., & Newman, C. B. (2016). Statin tolerability: In defence of placebo-controlled trials. *European journal of preventive cardiology*, 23(8), 891-896.

Whitley, E., Ball, J. (2002). Statistics review 5: Comparison of means. *Critical Care* 6:424-428.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution*, *3*(1), 129-137.

CHAPTER OVERVIEW

13: Assumptions of Parametric Tests

Introduction

Chapters 8, 10 and 12 were concerned primarily with tests of means among groups of treatments. ANOVA, t-tests, linear models, all involve estimation of parameters, qualities of populations. Although we have included assumptions about these statistics along the way, this chapter provides a summary about assumptions needed to be met in order to correctly interpret results of these statistical tests. Assumptions of parametric tests include how the data are presumed to be distributed (e.g., normality) and about the variability within groups (e.g., we assume equal variances). One important caveat: you can always estimate regardless of whether or not the assumptions are met. And, certainly, R and other statistical software will allow you to perform these calculations without warning. However, to the extent one or more assumptions do not hold, your conclusions, e.g., p-value and Type I error, will be influenced. That's what we mean by **statistical thinking** — knowing when your conclusions are valid.

This is the classic approach — provide tests of assumptions to justify use of ANOVA, etc. The modern approach, perhaps even the best practice approach, is instead to use more powerful statistical modeling approach, e.g., **generalized linear model (GLS)** to model for correlations among residuals (lack of independence assumption) or heteroscedastic variances (equal residual variances).

13.1: ANOVA assumptions

Introduction

Like all parametric tests, assumptions are made about the data in order to justify and trust estimates and inferences drawn from ANOVA. These are

1. Data come from **normal distributed population**. View with a histogram or Q-Q plot. Test with Shapiro-Wilks or other appropriate goodness of fit test†. Normality tests are the subject of Chapter 13.3.
2. **Sample size equal among groups**.
   - This is an example of a potentially **confounding factor** — If sample sizes differ, then any difference in means could be simply because of differences in sample size! This gets us into weighed versus unweighted means.
   - You shouldn't be surprised that modern implementations of ANOVA in software easily handle (adjust for) these known confounding factors. Depending on the program, you'll see "Adjusted means," "Least squares means," "Marginal means," etc. This just implies that the group means are compared after accounting for confounding factors.
   - Importantly, as long as sample sizes among the groups are roughly equivalent, normality assumption is not a big deal (low impact on risk of type I error).
3. **Independence of errors**. One consequence of this assumption is that you would not view 100 repeated observations of a trait on the same subject as 100 independent data points. We'll return to this concept more in the next two lectures. Some examples:
   - Colorimetric assay where the signal changes over time, and you measure in order (e.g., samples from group 1 first, samples from group 2 second, etc.) — this confounds group with time.
     - The consequence is that you are far more likely to reject the null hypothesis, committing a Type I error.
   - Let's say you are observing running speeds of ten mongoose. However, it turns out that five of your subjects are actually from the same family, identical quintuplets! Do you really have ten subjects?
   - Compare brain-body mass ratio among different species; this is a classic comparative method problem (Fig. 13.1.3). Since 1985 (Felsenstein 1985), it was recognized that the hierarchical evolutionary relationships among the species must be accounted for to control for lack of independence among the taxa tested. See Phylogenetically independent contrasts, Chapter 20.12.
4. **Equal variances among groups**. See Chapter 13.4 for how to test this for multiple groups.

Impact of assumptions

Note that R (and pretty much all statistics packages) will calculate the ANOVA and the p-value, but it is up to you to recognize that the P-value is accurate only if the assumptions are met. Violation of the assumption of normality can lead to Type I errors occurring more often than the 5% level. What to do if the assumptions are violated?

If the violation is due to only a handful of the data, you might proceed anyway. But following a significant test for normality, we could avoid the ANOVA in favor of nonparametric alternatives (Chapter 15), or, we might try to **transform the data**.

Consider a histogram of body-mass measures in grams for a variety of mammals (Fig. 13.1.1).

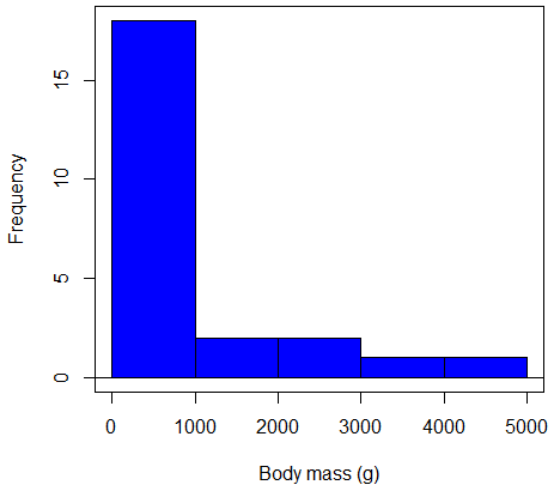


Figure 13.1.1: Histogram of body mass (g) for 24 mammals (data from Boddy et al 2012).

We will introduce a variety of statistical tests of the assumption of normality in Chapter 13.3, but looking at a histogram as part of our data exploration, we clearly see the data are right-skewed (Fig. 13.1.1). Is this an example of normal distributed sample of observations? Clearly not. If we proceed with statistical tests on the raw data set, then we are more likely to commit a Type I error (i.e., we will reject the null hypothesis more often than we should).

A note on normality and biology. It is VERY possible that data may not be normally distributed or have equal variances on the original scale, but a simple mathematical manipulation may take care of that. In fact, in many cases in biology that involve growth, many types of variables are expected to not be normal on the original scale. For example, while the relationship between body mass, M, and metabolic rate, MR, in many groups of organisms is allometric and increases positively, the relationship

$$MR = a \cdot M^b$$

is not directly proportional (linear) on the original scale. By taking the logarithm of both body mass and metabolic rate, however, the relationship is linear:

$$\log(MR) = \log(a) + b \cdot \log(M)$$

In fact, taking the logarithm (base 10, base 2, or base e) is often a common solution to both non-normal data (Fig. 13.1.2) and unequal variances.

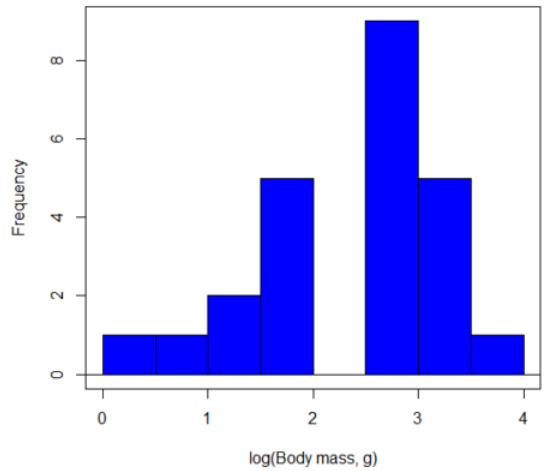


Figure 13.1.2: Histogram of $\log_{10}$-transformed body mass observations from Figure 13.1.1.

Other common transformations include taking the **square root** or the **inverse of the square-root** for skewed or kurtotic sample distributions, and the **arcsine** for frequencies (since frequencies can only be from 0 to 1 — need to "stretch the data" to make frequencies fit procedures like ANOVA). There are many issues about data transformation, but keep in mind three points. After completing the transformation, you should check the assumptions (normality, equal variances) again.

You may need to **recode the data before applying a transform**. For example, you cannot take the square root or logarithm of negative numbers. If you do not recode the data, then you will lose these observations in your subsequent analyses. In many cases, this problem is easily solved by adding 1 to all data points before making the transform. I prefer to make the minimum value 1 and go from there. The justification for data transformation is basically to realize that there is no necessity to use the common arithmetic or linear scale: many relationships are multiplicative or nonadditive (e.g., rates in biology and medicine).

Statistical outlier

Another topic we should at least mention here is the concept of outliers. While most observations tend to cluster around the mean or middle of the sample distribution, occasionally one or more observations may differ substantially from the others. Such values are called outliers, and we note that there are two possible explanations for an outlier:

1. the value could be an error.
2. it is a true value (and there may be an interesting biological explanation for its cause).

We encountered a clear outlier in the BMI homework. If the reason is (1), then we go back and either fix the error or delete it from the worksheet. If (2), however, then we have no objective reason to exclude the point from our analyses.

We worry if the outlier influences our conclusions — so it is a good idea to run your analyses with and without the outlier. If your conclusions remain the same, then no worries. If your conclusions change based on one observation, then this is problematic. For the most part you are then obligated to include the outlier and the more conservative interpretation of your statistical tests.

ANOVA is robust to modest deviations from assumptions

A comment about ANOVA assumptions ANOVA turns out to be robust to violations of item (1) or (2). That means unless the data are really skewed or the group sizes are very different, ANOVA will perform well (Type I error rate stays close to the specified 5% level). We worry about this however when p-value is very close to alpha!!

The third assumption is more important in ANOVA.

Like the t-test, ANOVA makes the assumption of equal variances among the groups, so it will be helpful to review why this assumption is important to both the t-test and ANOVA. In the two-sample independent t-test, the pooled sample variance, s_p^2, is taken as an estimate of the population variance, σ^2. If you recall,

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2} \longrightarrow s_p^2 = \frac{\sum_{i=1}^{2}\left[\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2\right]}{\sum_{i=1}^{2}\left(n_i - 1\right)}$$

where SS_1 refers to the sum of squares for the first group and SS_2 refers to the second group sum of squares (see our discussion on measures of dispersion) and v_1 refers to the degrees of freedom for the first group and v_2 refers to the second group degrees of freedom. We make a similar assumption in ANOVA. We assume that the variances for each sample are the same and therefore that they all estimate the population variance σ^2. To say it in another way, we are assuming that all of our samples have identical variability.

Once we make this assumption, we may pool (or combine) all of the SS's and DF's for all groups as our best estimate of the population variance, σ^2. The trick to understanding ANOVA is to realize that there can be two types of variability: there is variability due to being part of a group (e.g., even though ten human subjects receive the same calorie-restricted diet, not all ten will loose the same amount of weight) and there is variability among or between groups (e.g., on average, all subjects who received the calorie-restricted diet lost more weight than did those subjects who were on the non-restricted diet).

Example

The encephalization index (or encephalization quotient) is defined as the ratio of size the brain compared to body size. While there is a well-recognized increase in brain size given increased body size, encephalization describes a shift of function to cortex (frontal, occipital, parietal, temporal) from noncortical parts of the brain (cerebellum, brainstem). Increased cortex is associated with increased complexity of brain function; for some researchers, the index is taken as a crude estimate of intelligence. Figure $13.1.3A$ shows plot of brain mass in grams versus body size (grams) for 24 mammal species (data sampled from Boddy et al 2012); figure $13.1.3B$ shows the same data, but following $\log_{10}$-transform of both variables.

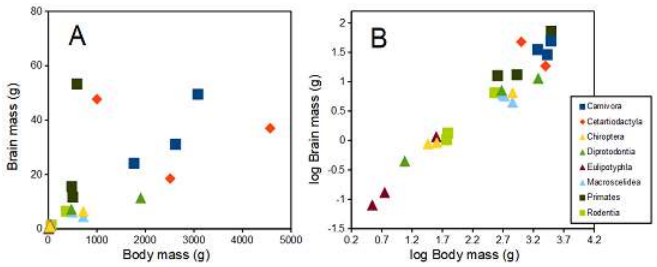


Figure 13.1.3: Plot of brain and body weights (A) and $\log_{10}$-$\log_{10}$ transform (B) for a variety of species (data from Boddy et al 2012). The ratio is called encephalization index.

Looking at the two figures, the linear relationship between the two variables is obvious in Figure 13.1.3*B*, less so for Figure 13.1.3*A*. Thus, one biological justification for transformation of the raw data is exemplified with the brain-body mass dataset: the association is allometric, not additive. The other reason to apply a transform is statistical; the $\log_{10}$-transform improves the normality of the variables. Take a look at the Q-Q plot for the raw data (Figure 13.1.4) and for the log `10` -transformed data (Fig. 13.1.5).

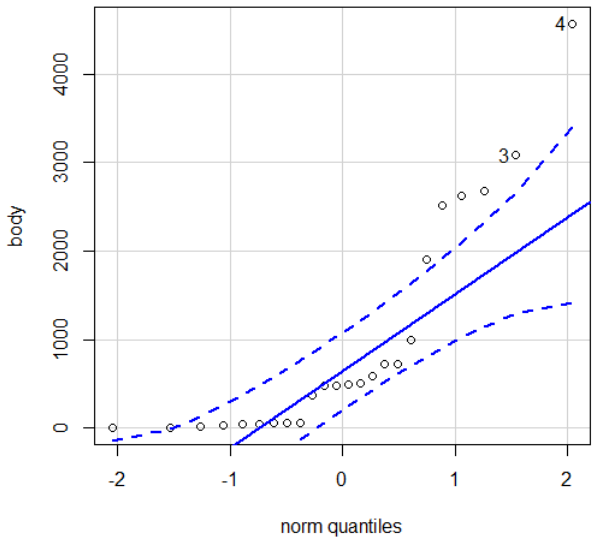


Figure 13.1.4: Q-Q plot, raw data. Compare to Figure 13.1.1.

Note the data don't fall on the straight line; a few fall outside of the confidence interval (the curved dashed lines), which suggests the data are not normally distributed (see histogram, Figure 13.1.1). And for the transformed data, the Q-Q plot is shown in Fig. 13.1.5.

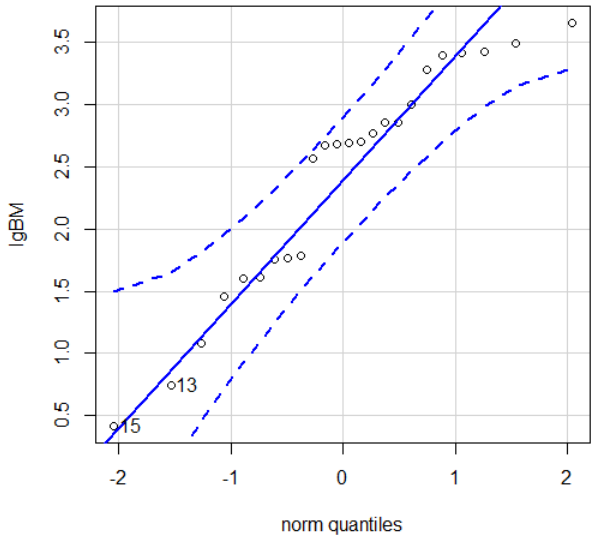


Figure 13.1.5: Q-Q plot of same data, $\log_{10}$-transformed. Compare to Figure 13.1.2.

Compared to the raw data, the transformed data now fall on the line and none are outside of the confidence interval. We would conclude that the transformed data are more normal, thus, better meeting the assumptions of our parametric tests.

Lack of independence among data

Species comparisons are common in evolutionary biology and related fields. As noted earlier, comparative data should not be treated as independent data points. For our 24 species, I plotted the estimate of the phylogeny (timetree.org).

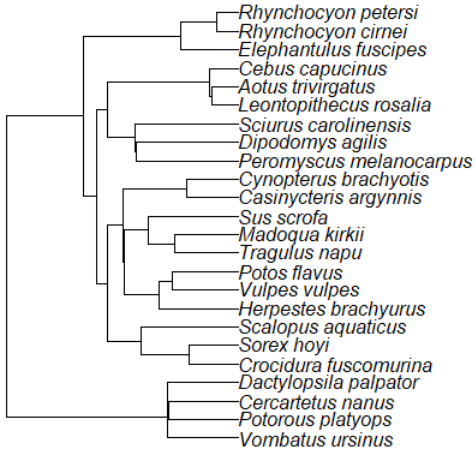


Figure 13.1.6: Phylogenetic tree of 24 species used in this report.

The conclusion? We don't have 24 data points, more like 8 points. Because the species are more or less related, there are fewer than 24 independent data points. Statistically, this would mean that the errors are correlated. Various approaches to account for this lack of independence have been developed; perhaps the most common approach is to apply phylogenetically independent contrasts, a topic discussed in Chapter 20.12. (Boddy et al 2012 used this approach.)

Note:

See Chapter 20.11 for help making a plot like the one shown in Figure 13.1.6.

Questions

1. †Shapiro-Wilks is one test of normality. Can you recall the name of the other normality test we named?

Data set and R code used in this page

```
species, order, body, brain
'Herpestes ichneumon', Carnivora, 1764, 24.1
'Potos flavus', Carnivora, 2620, 31.05
'Vulpes vulpes', Carnivora, 3080, 49.5
'Madoqua kirkii', Cetartiodactyla, 4570, 37
'Sus scrofa', Cetartiodactyla, 1000, 47.7
'Tragulus napu', Cetartiodactyla, 2510, 18.5
'Casinycteris argynnis', Chiroptera, 40.5, 0.92
'Cynopterus brachyotis', Chiroptera, 29, 0.88
'Potorous platyops', Chiroptera, 718, 6.5
'Cercartetus nanus', Diprotodontia, 12, 0.44548
'Dactylopsila palpator', Diprotodontia, 474, 7.15876
'Vombatus ursinus', Diprotodontia, 1902, 11.396
'Crocidura fuscomurina', Eulipotyphla, 5.6, 0.13
'Scalopus aquaticus', Eulipotyphla, 39.6, 1.48
'Sorex hoyi', Eulipotyphla, 2.6, 0.107
'Elephantulus fuscipes', Macroscelidea, 57, 1.33
'Rhynchocyon cirnei', Macroscelidea, 490, 6.1
'Rhynchocyon petersi', Macroscelidea, 717.3, 4.46
'Aotus trivirgatus', Primates, 480, 15.5
'Cebus capucinus', Primates, 590, 53.28
'Leontopithecus rosalia', Primates, 502.5, 11.7
'Dipodomys agilis', Rodentia, 61.4, 1.34
'Peromyscus melanocarpus', Rodentia, 58.8, 1.03
'Sciurus carolinensis', Rodentia, 367, 6.49
```

The Newick code for the tree in Figure 13.1.6.

```
((Vombatus_ursinus:48.94499077,
((Potorous_platyops:47.59556667,Cercartetus_nanus:47.59556667)'14':0.66887333,Dactylo
(((((Crocidura_fuscomurina:33.74066667,Sorex_hoyi:33.74066667)'10':33.03022424,Scalop
(((Herpestes_brachyurus:54.32144118,
(Vulpes_vulpes:45.52834967,Potos_flavus:45.52834967)'9':8.79309151)'22':23.43351523,
((Tragulus_napu:43.96862857,Madoqua_kirkii:43.96862857)'8':17.99735995,Sus_scrofa:61.
(Casinycteris_argynnis:35.20000000,Cynopterus_brachyotis:35.20000000)'29':43.32874208
(((Peromyscus_melanocarpus:69.89837667,Dipodomys_agilis:69.89837667)'43':0.64655123,S
((Leontopithecus_rosalia:18.38385647,Aotus_trivirgatus:18.38385647)'40':1.29720005,Cel
(Elephantulus_fuscipes:39.23366667,
(Rhynchocyon_cirnei:15.34500000,Rhynchocyon_petersi:15.34500000)'39':23.88866667)'56'
```

13.2: Why tests of assumption are important

Introduction

Note that R (and pretty much all statistics packages) will calculate t-tests or ANOVA or whatever test you ask for, and return a p-value, but it is up to you to recognize that the p-value is accurate only if the assumptions are met. Thus, you can always estimate a parameter, but interpret with caution. The great thing about statistics is that you can directly evaluate whether assumptions hold.

Violation of the assumptions of normality or equal variances can lead to Type I errors occurring more often than the 5% level. That means you will reject the null hypothesis more often than you should! If the goal of conducting statistical tests on results of an experiment is to provide confidence in your conclusions, then failing to verify assumptions of the test are no less important than designing the experiment correctly in the first place. Thus, the simple rule is: know your assumptions, test your assumptions. Evaluating assumptions is a learned skill:

- conduct proper data exploration
- use specialized statistical tests to evaluate assumptions
  - data normally distributed? e.g., histogram, Shapiro-Wilk test
  - groups equal variance? e.g., box-plot, Levene's median test
- evaluate influence of any outlier observations

What to do if the assumptions are violated?

This is where judgement and critical thinking apply. You have several options. First, if the violation is due to only a few observations, then you might proceed anyway, in effect invoking the **Central Limit Theorem** as justification. Second, you could check your conclusions with and without the few observations that seem to depart from the trend in the rest of the data set — if your conclusion holds without the "outliers", then you might conclude that your results are robust). Third, you might apply a **data transform**, reasoning that the distribution from which the data were sampled was log-normal, for example. Applying a log transform (natural log, base 10, etc.,) will tend to make the variances less different among the groups and may also improve normality of the samples. Fourth, if a nonparametric test is available, you might use it instead of the parametric test. For example, we discuss the Levene's test of equal variances as a better choice than the parametric F-test. Additionally, there are many nonparametric alternatives to parametric tests. For example, t-tests are parametric and their non-parametric alternatives include tests like Wilcoxon and Mann-Whitney. ANOVA, too is parametric, and its nonparametric alternative version is called **rank ANOVA** (see also Kruskal-Wallis). See Chapter 15 for nonparametric tests. Finally, a resampling approach could be taken, where the data themselves are used to generate all possible outcomes like the Fisher Exact test; with large sample size, bootstrap or Jackknife procedures are used (Chapter 19).

For now, let's introduce you to the kinds of nonparametric statistical tests for which the t-test is just one example. For the independent sample t-test, our first method to account for the possible violation of equal variances is a parametric test, Welch's variation of the Student's t-test. Instead of the pooled standard deviation, Welch's test accounts for each group's variance in the denominator.

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom for the Welch's test are now

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot v_1} + \frac{s_2^4}{n_2^2}}$$

where df for Student's t-test was $n_1 + n_2 - 2$. Note that Welch's test requires normal distributed data. Note also that in R Commander, the default option for conducting the t-test is Welch's version, not the standard t-test (Fig. 13.2.1).

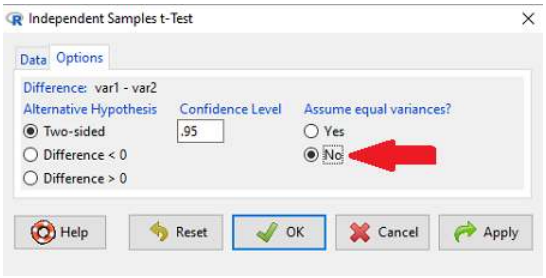


Figure 13.2.1: Copy and Paste Caption here. (Copyright; author via source)

See discussions in Olsen (2003), Choi (2005), and Hoekstra et al (2012).

Questions

Please read article:

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in psychology*, *3*, 137. Link to article

From the article and your readings in Mike's Biostatistics Book, answer the following questions:

1. What are some consequences if a researcher fails to check statistical assumptions?
2. Explain why use of graphics techniques may be more important than results of statistical tests for checks of statistical assumptions.
3. Briefly describe graphical and statistical tests of assumptions of (1) normality and (2) equal variances
4. Pick a biological research journal for which you have online access to recent articles. Pick ten articles that used statistics (e.g., look for "t-test" or "ANOVA" or "regression"; exclude meta analysis and review articles — stick to primary research articles). Scan the Methods section and note whether or not you found a statement that confirms if the author(s) checked for violation of (1) normality and (2) equal variances. Construct a table.
5. Review your results from question 3. Out of the ten articles, how many reported assumption checking? How does your result compare to those of Hoekstra et a; (2012)?

13.3: Test assumption of normality

Introduction

I've commented numerous times that your conclusions from statistical inference are only as good as the validity of making and applying the correct procedures. This implies that we know the assumptions that go into the various statistical tests, and where possible, we critically test the assumptions. From time to time, then, I will provide you with "tests of assumptions."

Here's one. The assumption of normality, that your data were sampled from a population with a normal distribution for the variable of interest, is key and there are a number of ways to test this assumption. Hopefully as you read through the next section you can extend the logic to any distribution; if the data are presumed to come from a binomial, or a Poisson, or a uniform distribution, then the logic of goodness of fit tests would apply.

How to test normality assumption

It's not a statistical test *per se*, but the best option is to simply plot (histogram) the data. You can do these graphics by hand, or, install the data mining package `rattle` which will generate nice plots useful for diagnosing normality issues.

> **Note:**
>
> `rattle` (R Analytic Tool To Learn Easily) is a great "data-mining" package. Rattle (version 5.5.1 as of March 2023) provides a graphical user interface which makes it straightforward to work with. It doesn't work well with `Rcmdr` , but can be used along with `RStudio` . You'll need to also install `RGtk2` and `cairoDevice` packages, which, unfortunately, were removed from CRAN in late 2021. Therefore, if you wish to run rattle you'll need to install older versions of `RGtk2` and `cairoDevice` — as of March 2023, follow instructions for your operating system at https://rattle.togaware.com/.

The rattle histogram plot superimposes a normal curve over the data, which allows you to "eyeball" the data.

First, the eye test. I used the R-package `rattle` for this on a data set of comet tail lengths of rat lung cells exposed to different amounts of copper in growth media (scroll to bottom of page or click here to get the data).

In addition to the histogram (Fig. 13.3.1 top image), I plotted the cumulative function (Fig. 13.3.1 bottom image). In short, if the data set were normal, then the cumulative frequency plot should look like a straight line.

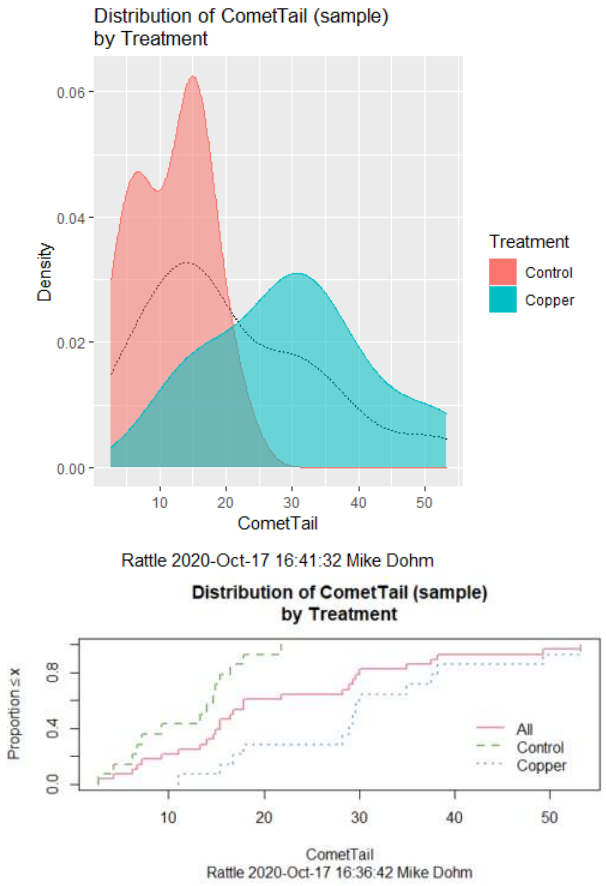


Figure 13.3.1: `rattle` descriptive graphics on Comet Copper dataset. Dotted line (top image) and red line (bottom image) follow the combined observations regardless of treatment.

So, just looking at the data set, we don't see clear evidence for a normal-like data set. The top image (Fig. 13.3.1) looks stacked to the left and the cumulative plot (bottom image) is bumped in the middle, not falling on a straight line. We'll need to investigate the assumption of normality more for this data set. We'll begin by discussing some hypothetical points first, then return to the data set.

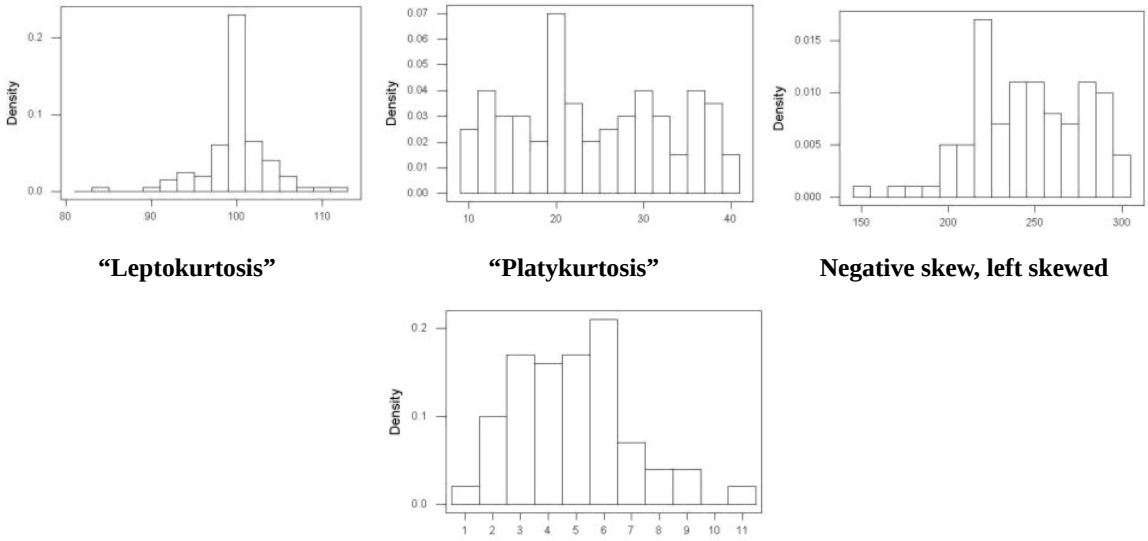
Goodness of fit tests of normality hypothesis

While graphics and the "eye ball test" are very helpful, you should understand that whether or not your data fits a normal distribution; that's a testable hypothesis. The null hypothesis is that your sample distribution fits are normal distribution. In general terms, these "fit" hypotheses are viewed as "goodness of fit" tests. Often times, the test is some variation of a χ^2 problem: you have your observed (sample distribution) and you compare it to some theoretical expected value (e.g., in this situation, the normal distribution). If your data fit a normal curve, then the test statistics will be close to zero.

We have discussed before that the data should be from a normal distributed population. To the extent this is true, then we can trust our statistical tests are performing the way they are expected to do. We can test the assumption of normality by comparing our sample distribution against a theoretical distribution, the normal curve. I've shown several graphs in the past that "looked normal". What are the alternatives for **unimodal** (single peak) distributions?

Kurtosis describes the shape of the distribution, whether it is stacked up in the middle (leptokurtosis), or more spread out and flattened (platykurtosis).

Skewness describes differences from symmetry about the middle. For example, left skew means the tail of the distribution extends to the left, i.e., smaller values are more prevalent than larger values.



“Leptokurtosis” **“Platykurtosis”** **Negative skew, left skewed**

Positive skew, right skewed

Figure 13.3.1: Graphs describing different distributions. From top to bottom: Leptokurtosis, platykurtosis, negative skew, positive skew.

The easiest procedures for goodness of fit tests of normality are based on the χ^2 distribution and yield a “goodness of fit” test for normal distribution. We discussed the χ^2 distribution in Chapter 6.9, and used the test in Chapter 9.1.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O refers to the observed data (what we’ve got) and E refers to the expected (e.g., data from a normal curve with same mean and standard deviation as our sample).

To illustrate, I simulated a data set in R.

Rcmdr: Distributions → Continuous distributions → Normal distribution → Sample from normal distribution

I created 100 values, stored them in column 1, and set the population mean = 125 and population standard deviation = 10. And therefore the population standard error of the mean was 1.0.

The resulting MIN/MAX was from 99.558 to 146.16; the sample mean was 124.59 with a sample standard deviation of 9.9164. And therefore the sample standard error of the mean was 0.9916.

Question: After completing the steps above, your data will be slightly different from mine… Why?

But getting back to the main concept, does our data agree with a normal curve? We have discussed how to construct histograms and frequency distributions.

Let’s try six categories (why six? we discussed this when we talked about histograms).

All chi-square tests are based on categorical data, so we use the counts per category to get our data. Group the data, then count the number of OBSERVED in each group. To get the EXPECTED values, use the Z-score (normal deviate) with population mean and standard deviation as entered above.

Table 13.3.1. Tabulated values for test of normality.

Number of observations	Weight	Normal deviate (Z)	Expected Proportion	Expected number	(Obs – Exp)^2 / Exp
105	3	less than or equal to -2	0.0228	2.28	0.227368421
105 < 115	17	between -1 & -2 = 0.1587 – 0.0228	0.1359	13.59	0.855636497

Number of observations	Weight	Normal deviate (Z)	Expected Proportion	Expected number	(Obs – Exp)^2 / Exp
115 < 125	34	between -0 & -1 = 0.5 – 0.1587	0.3413	34.13	0.000495166
125 < 135	30	between +0 & +1 = 0.5 – 0.1587	0.3413	34.13	0.499762672
135 < 145	15	between +1 & +2 = 0.9772 – 0.8413	0.1359	13.59	0.146291391
> 145	1	greater than or equal to +2 = 1 – 0.9772	0.0228	2.28	0.718596491
					$\chi^2 = 2.44815064$

Then obtain the critical value of the χ^2 with $df = 6 - 1 = 5$ (see Appendix A.3, Table of chi-square critical values , critical $\chi^2 = 11.1$ with $df = 5$).

Thus, we would not reject the null hypothesis and would proceed with the assumption that our data could have come from a normally distributed population.

This would be an OK test, but different approaches, although based on a chi-square-like goodness of fit, have been introduced and are generally preferred. We have just shown how one could use the chi-square goodness of fit approach to testing whether your data fit a normal distribution. A number of modified tests based on this procedure have been proposed; we have already introduced and used one (Wilks-Shapiro), which is easily accessed in R.

Rcmdr: Summaries → Wilks-Shapiro

Like Wilks-Shapiro, another common "goodness-of-fit" test of normality is the **Anderson-Darling test**. This test is now included with `Rcmdr` , but it's also available in the package `nortest` . After the package is installed and you run the library (you should by now be able to do this!), then at the R prompt (>) type:

```
ad.test(dataset$variable)
```

replacing " `dataset$variable` " with the name of your data set and variable name.

In the context of goodness of fit, a perfect fit means the data are exactly distributed as a normal curve, and the test statistics would be zero. Differences away from normality increase the value of the test statistic.

How do these tests perform on data?

The histogram of our simulated normal data of 100 observations with mean = 125 and standard deviation = 10 is shown in Fig. 13.3.3

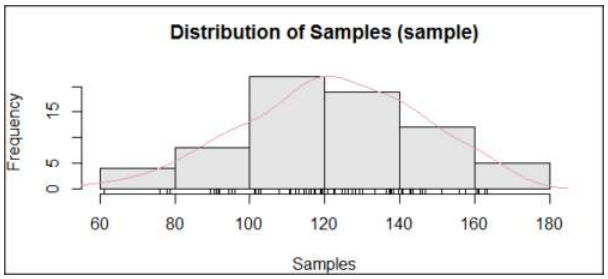


Figure 13.3.3: Histogram of simulated normal dataset, $\mu = 125$, $\sigma = 10$.

and here's the cumulative plot (Fig. 13.3.4).

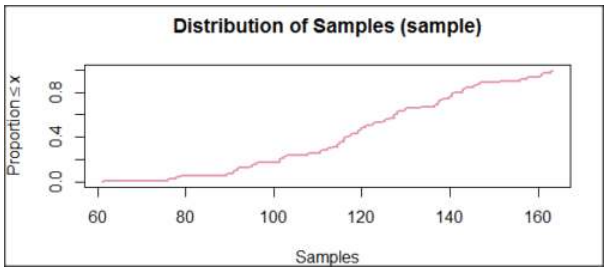


Figure 13.3.4: Cumulated frequency plot of simulated normal dataset, $\mu = 125$, $\sigma = 10$.

Results from Anderson-Darling test were $A = 0.2491$, p-value = 0.7412, where A is the Anderson-Darling test statistic.

Results of the Shapiro-Wilks test on the same data: $W = 0.9927$, p-value = 0.8716, where W is the Shapiro-Wilks test statistic. We would not reject the null hypothesis in either case because $p > 0.05$.

Example

Histogram of a data set, highly skewed to the right. 90 observations in three groups of 30 each, with mean = 0 and standard deviation = 1 (Fig. 13.3.5)

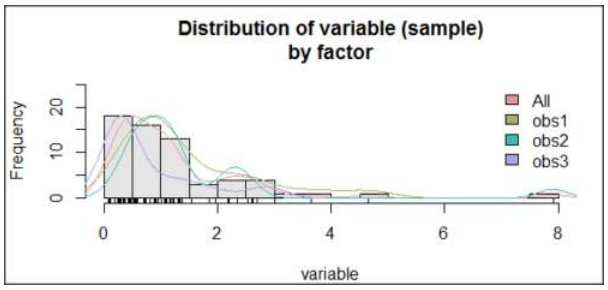


Figure 13.3.5: Histogram of simulated normal dataset, $\mu = 0$, $\sigma = 1$.

and the cumulative frequency plot (Fig. 13.3.6)

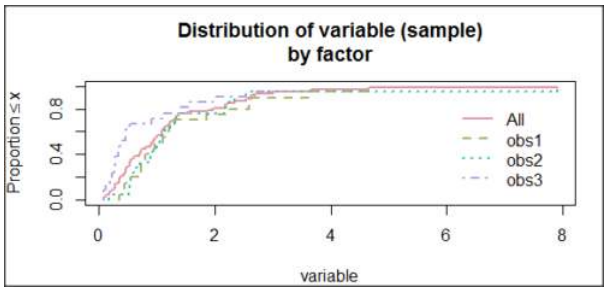


Figure 13.3.6: Cumulated frequency of simulated normal dataset, $mu = 0$, $\sigma = 1$.

Results from Anderson-Darling test were $A = 9.0662$, p-value $< 2.2 \times 10^{-16}$. Results of the Shapiro-Wilks test on the same data were $W = 0.6192$, p-value = 6.248×10^{-14}. Therefore, we would reject the null hypothesis because $p < 0.05$.

The Shapiro-Wilk test in Rcmdr

Let's go back to our data set and try tests of normality on the entire data set, i.e., not by treatment groups.

Rcmdr: Statistics → Summaries → Test of normality…

```
normalityTest(~CometTail, test="shapiro.test", data=CometCopper)

 Shapiro-Wilk normality test

data: CometCopper$CometTail
W = 0.91662, p-value = 0.006038
```

End of R output

Another test, built on the basic idea of a chi-square goodness of fit, is the Anderson-Darling test. Some statisticians prefer this test and it is one built into some commercial statistical packages (e.g., Minitab). To obtain the Anderson-Darling test in R, you need to install a package. After installing `nortest` package, run the AD test at the command prompt.

```
require(nortest)
ad.test(CometCopper$CometTail)

Anderson-Darling normality test

data: CometCopper$CometTail
A = 1.0833, p-value = 0.006787
```

End of R output

The *p*-values are both much less than 0.05, so we would reject the assumption of normality for this data set.

Which test of normality?

Why show you two tests for normality, the Shapiro-Wilks and Anderson-Darling? The simple answer is that both are good as general tests of normality, both are widely used in scientific papers, so just pick one and go with it as your general test of normality.

The more complicated answer is that each is designed to be sensitive to different kinds of departure from normality. By some measures, the Shapiro-Wilks test is somewhat better (i.e., has more statistical power to test the null hypothesis) than other tests, but this is not something you want to get into as a beginner. So, I show both of them to you so that you are at least introduced to the concept that there is often more than one way to test a hypothesis. The bottom line is that plots may be best!

Questions

1. Work describe in this chapter involves statistical tests of the assumption of normality. It is just as important, maybe more so, to also apply graphics to take advantage of our built-in pattern recognition functions. What graphic techniques, besides histogram, should be used to view the distribution of the data?
2. In R, what command would you use so that you can call the variable name, `CometTail` , directly instead of having to refer to the variable as `CometCopper$CometTail` ?
3. Why are Anderson-Darling, Shapiro-Wilks and other related tests referred to as “goodness of fit” tests? You may wish to review discussion in Chapter 9.1.
4. The example tests presented for the Comet Copper data set were conducted on the whole set, not by treatment groups. Re-run tests of normality via `Rcmdr` , but this time, select the By groups option and select Treatment.

Data set used in this page

Treatment	CometTail
Control	17.86
Control	16.52
Control	14.93
Control	14.03
Control	13.33
Control	8.81
Control	14.70
Control	9.26
Control	21.78

Treatment	CometTail
Control	6.18
Control	9.20
Control	5.54
Control	6.72
Control	2.63
Control	7.19
Control	5.39
Control	11.29
Control	15.44
Control	17.86
Contro	l4.25
Copper	53.21
Copper	38.93
Copper	18.93
Copper	30.00
Copper	28.93
Copper	15.36
Copper	17.86
Copper	17.50
Copper	21.07
Copper	29.29
Copper	28.21
Copper	16.79
Copper	21.07
Copper	37.50
Copper	38.22
Copper	17.86
Copper	29.64
Copper	11.07
Copper	35.00
Copper	49.29

13.4: Tests for equal variances

Introduction

In order to carry out statistical tests correctly, we must test our data first to see if our sample conforms to the assumptions of the statistical test. If our data do not meet these assumptions, then inferences drawn may be incorrect. How far off our inferences may be depends on a number of factors, but mostly it depends on how far from the expectations our data are.

One assumption we make with parametric tests involving ratio scale data is that the data could be from a normally distributed population. The other key assumption introduced, but not described in detail for the two-sample t-test, was that the variability in the two groups must be the same, i.e., **homoscedasticity**. Thus, in order to carry out the independent sample t-test, we must assume that the **variances are equal**.

There are two general reasons we may want to concern ourselves with a test for the differences between two variances:

1. The t-test (and other tests like one-way ANOVA) requires that the two samples compared have the same variances. If the Variances are Not Equal we need to perform a modified t-test (see Welch's formula).
2. We may also be interested in the differences between the variances in two populations.

Example 1: In genetics we might be interested in the difference between the variability of response of inbred lines (little genetic variation but environmental variation) versus an outbred population (lots of genetic and environmental variation).

Example 2: Environmental stress can cause organisms to have developmental instability. This might cause organisms to be more variable in morphology, or the two sides (right & left) of an organism may develop non-symmetrically. Therefore, polluted environments might cause organisms to have greater variability compared to non-polluted environments.

The first way to test the variances is to use the F-test. This works for two groups.

$$F = \frac{s_1^2}{s_2^2}$$

For more than two groups, we'll use different tests (e.g., Bartlett's test, Levene's test).

Remember that the formula for the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)}{n-1}$$

The Null Hypothesis is that the two samples have the same variances: $H_O : s_1^2 = s_2^2$

The Alternate Hypothesis is that the two samples do not have the same variances: $H_A : s_1^2 \neq s_2^2$

Note: I prefer to evaluate this as a one-tailed test: identify the larger of the two variances and take that as the numerator Then, the null hypothesis is $H_O : s_1^2 \leq s_2^2$

and therefore, the alternative hypothesis is $H_A : s_1^2 > s_2^2$ (i.e., a one-tailed test).

Another way to state equal variance test is that we are testing for **homogeneity of variances**. You may run across the term **homoscedasticity**; it is the same thing, just a different three-dollar word for "equal variances."

Stated yet another way, if we reject the null hypothesis, then the variances are unequal or show **heterogeneity**. An additional and equivalent three-dollar word for inequality of variances is called **heteroscedasticity**.

More about the F-test

For the F-test, the null hypothesis is that the variances are equal. This means that the "expected" F value will be one: $F = 1.0$. (The ***F*-distribution** differs from t-distribution because it requires 2 values for DF, and ranges from 1 to infinity for every possible combination of v_1 and v_2).

To evaluate the null hypothesis we need the **degrees of freedom**. For the F-test we need two different degrees of freedom, one set for each group): from the table in *Appendix A.3 – F distribution*, look up 5% Type I error line in this table because we make it one-tailed.

I need the F-test statistic at $F_{0.05_1, v_1, v_2}$.

Examples of difference between two variances, Table 13.4.1.

Sample 1: Aggressiveness of Inbred Mice (number of bites in 30 minutes)

Sample 2: Aggressiveness of Outbred Mice (number of bites in 30 minutes)

Table 13.4.1. Aggression by inbred and outbred mice.

Sample 1 Aggressiveness of Inbred Mice (number of bites in 30 minutes)	**Sample 2** Aggressiveness of Outbred Mice (number of bites in 30 minutes)
3	4
5	10
4	4
3	7
4	7
5	10
4 = mean	7 = mean

1. Identify the null and alternate hypotheses
2. Calculate variances
3. Calculate F-test
4. The "test statistic" for this hypothesis test was $F = 7.2/0.8 = 9.0$
5. Determine Critical Value of the F table (*Appendix – F distribution table*)

Example of how to find the critical values of the F distribution for $F_{0.05_2}$, numerator $DF_{v_1} = 5$ and denominator $DF_{v_2} = 5$.

Table 13.4.2. Portion of F distribution, see Appendix – F distribution.

$\alpha = 0.05$		v_1							
		1	2	3	4	5	6	7	8
v_2	1					230			
	2					19.3			
	3					9.01			
	4					6.26			
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
	6								

Or, instead of using tables, use R.

Rcmdr: Distributions → F distribution → F probabilities

and enter the numbers as shown below (Fig. 13.4.1).

Figure 13.4.1: Screenshot of F distribution probabilities in R Commander.

This will return the p-value, and you would interpret this against your Type I error rate of 5% as you do for other tests.

From Table 2, *Appendix – F distribution* we find $F_{0.05(1),5,5} = 5.05$

And the p-value = 0.015

```
pf(c(9), df1=5, df2=5, lower.tail=FALSE)
[1] 0.01537472
```

Question: Reject or Accept null hypothesis?

Question: What is the biological interpretation or conclusion from this result?

R code

Rather than play around with the tables of critical values, which are awkward and old-school (and I am showing you these stats tables so that you get a feel for the process, not so you'd actually use them in real practice), use Rcmdr to generate the F test and therefore return the F distribution probability value. As you may expect, R provides a number of options for testing the equal variances assumption, including the F test. The F test is limited to only two groups and, because it is a parametric test, it also makes the assumption of normality, so the F test should not be viewed as necessarily the best test for the equal variances assumption among groups. We present it here because it is a logical test to understand and because of its relevance to the Mean Square ratios in the ANOVA procedures.

So, without further justification, here is the presentation on how to get the F test in Rcmdr. At the end of this section I present a better procedure than the F test for evaluating the equal variance assumption called the Levene test.

Return to the bite data in the table above and enter the data into an R data frame. Note that the data in the table above are unstacked; R expects the data to be stacked, so either create a stacked worksheet and transcribe the data appropriately into the cells of the worksheet, or, go ahead and enter the values into two separate columns then use the Stack variables in active data set… command from the Data menu in Rcmdr.

Then, proceed to perform the F test.

Rcmdr: Statistics → Variances → Two variances F-test…

The first context menu popup is where you enter the variables (Fig. 13.4.2):

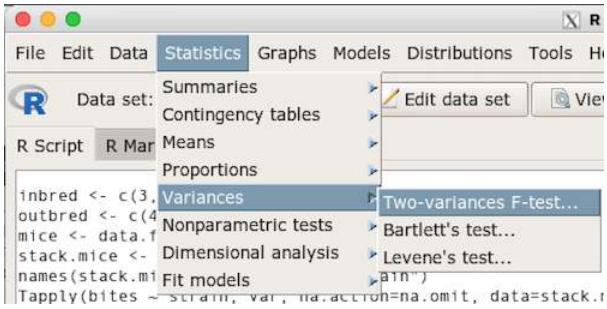


Figure 13.4.2: Screenshot of how to access to the Two-variances F-test in R Commander.

Because there are only two variables in the data set and because `Strain` contains the text labels of `inbred` or `outbred` whereas the other variable is numeric data type, R will correctly select the variables for you by default. Select the "Options" tab to set the parameters of the F-test (Fig. 13.4.3).

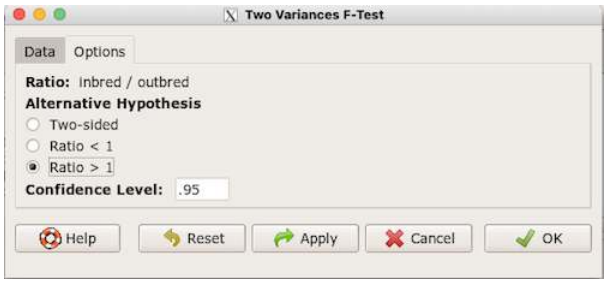


Figure 13.4.3: Screenshot of menu options for R Commander F test.

When you are finished setting the alternative hypothesis and **confidence levels**, proceed with the F-test by clicking the OK button.

```
    F test to compare two variances

data:  mice.aggression
F = 0.1111, num df = 5, denom df = 5, p-value = 0.03075
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.01554788 0.79404243
sample estimates:
ratio of variances
         0.1111111
```

End of R output

Levene's test of equal variances

We will discuss this test in more detail following our presentation on ANOVA. For now, we note that the test works on two or more groups and is a conservative test of the equal variance assumption. Nonparametric tests in general make fewer assumptions about the data and in particular make no assumption of normality like the F test. It is in this context that the Levene's test would be preferable over the F test. Below we present only how to calculate the statistic in R and Rcmdr and provide the output for the same mouse data set.

Assuming the data are stacked, obtain the Levene's test in Rcmdr by clicking on **Rcmdr: Statistics → Variances → Levene's test…** (Fig. 13.4.4)

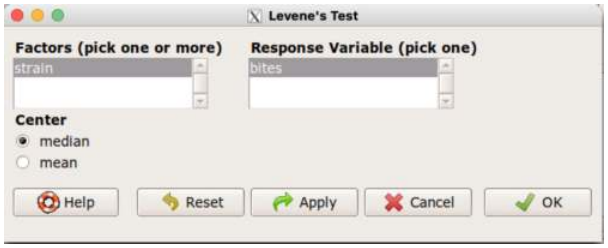


Figure 13.4.4: Screenshot of menu options for R Commander Levene's test.

Select the median and the factor variable (in our case "Strain") and the numeric outcome variable ("Bites"), then click OK button.

```
Tapply(bites ~ strain, var, na.action=na.omit, data=mice.aggression) # variances by g
inbred outbred
   0.8     7.2

leveneTest(bites ~ strain, data=mice.aggression, center="median")
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
```

```
group  1       4 0.07339 .
      10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

End of R output

Note that the p-values do not agree between the parametric F test and the nonparametric Levene's test! If we were to go by the results of the F test, the p-value was 0.031, less than our Type I error of 5%, and we would tentatively conclude that the assumption of equal variances may not apply. On the other hand, if we go with the Levene's test the p-value was 0.074, which is greater than our Type I error rate of 5% and we would therefore conclude the opposite, that the assumption of equal variances might apply! Both conclusions can't hold, so which test result of equal variances do we prefer, the parametric F test or the nonparametric Levene's test?

Cahoy's bootstrap method

Draft. Cahoy (2010). Variance-based statistic bootstrap test of heterogeneity of variances. We discuss bootstrap methods in Chapter 19.2; in brief, bootstrapping involves resampling the dataset and computing a statistic on each sample. This method may be more powerful, that is, more likely to correctly reject the null hypothesis when warranted, compared to Levene's test.

For now, install package `testequavar` .

Function for testing two samples:

```
equa2vartest(inbred, outbred, 0.05, 999)
```

R output:

```
[[1]]
[1] "Decision: Reject the Null"
$Alpha
[1] 0.05
$NumberOfBootSamples
[1] 999
$BootPvalue
[1] 0.006
```

The output "Decision: Reject the Null" reflects output from a box-type acceptance region.

Compare results from Levene's test: p-value 0.07339 suggests accept hypothesis of equal variances, whereas bootstrap method indicates variances heterogenous, i.e., reject equal variance hypothesis. However, re-running the test without setting the seed for R's pseudorandom number generator will result in different p-values. For example, I re-ran Cahoy's test five times with the following results:

```
0.008
0.006
0.002
0.01
0.004
```

Questions

1. Test assumption of equal variances by Bartlett's method and by Levene's test on `OliveMoment` variable from the Comet tea data set introduced in Chapter 12.1. Do the methods agree? If not, which test result would you choose?

- BONUS. Retest homogeneous variance hypothesis by `equa3vartest(Copper.Hazel, Copper, Hazel, 0.05, 999)` . Reject or fail to reject the null hypothesis by bootstrap method?

2. Test assumption of equal variances by Bartlett's method and by Levene's test on `Height` from the O'hia data set introduced in Chapter 12.7. Do the methods agree? If not, which result would you choose?
   - BONUS. Retest homogeneous variance hypothesis by `equa3vartest(M.1, M.2, M.3, 0.05, 999)` . Reject or fail to reject the null hypothesis by bootstrap method?

R code reminders

The bootstrap method expects the variables in unstacked format. A simple method to extract the variables from the stacked data is to use a command like the following. For example, extract OliveMoment values for Copper-Hazel treatment.

```
Copper.Hazel <- cometTea$OliveMoment[1:10]
```

For Copper, Hazel, replace above with [11:20], and [21-30] respectively. Note: changed variable name from Copper-Hazel to Copper.Hazel. The hyphen is a reserved character in R.

13.5: Chapter 13 References and Suggested Readings

Cahoy, D. O. (2010). A bootstrap test for equality of variances. *Computational statistics & data analysis*, *54*(10), 2306-2316.

Choi, P. T. (2005). Statistics for the reader: what to ask before believing the results. *Canadian Journal of Anesthesia*, 52(1), R46-R46.

Glass, G. V., Peckham, P. D., Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research* 42:237-288.

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)?. *Frontiers in psychology*, 3, 137.

Nakagawa, S., Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Review* 82:591–605.

Olsen, C. H. (2003). Review of the use of statistics in infection and immunity. *Infection and immunity*, 71(12), 6689-6692.

CHAPTER OVERVIEW

14: ANOVA Designs, Multiple Factors

Introduction

In our previous discussions about t-tests and ANOVA we focused on procedures with one dependent (response) variable and a single independent (predictor) factor variable that may cause variation in the response variable. In this chapter we extend our discussions about the **general linear model** by

1. Reviewing the one-way ANOVA, and providing a few examples of the one-way design.
2. Reviewing and setting the stage for adding a second independent variable to the model.

Additional one-way ANOVA examples

1. In a plants, we may have a response variable like height and one factor variable (location: sun vs. shade) thought to influence plant height (e.g. Aphalo et al 1999).
2. Pulmonary macrophage phagocytosis behavior (response variable) after exposure of toads to clean air or ozone (factor with 2 levels) (Dohm et al. 2005).
3. Monitor weight change on subjects after 6 weeks eating different diet (DASH, control) (Elmer et al. 2006).

All three of the examples are based on the same **statistical model** which may be written as:

$$Y_{ik} = \mu + A_i + \epsilon_{ik}$$

where μ is the grand mean, Y is the response variable and A is the independent variable, or factor, with $k = 1, 2, \ldots K$ levels, groups, or treatments. The total number of experimental units (e.g., subjects) is given by $i = 1, 2, 3, ldotsn$. Note that in the first and third examples, because there were only two groups (example 1: k = location, shade; example 3: k = DASH, control). Note that this problem could have been evaluated as an independent sample t-test. For the second example, there were three groups so k = clean air, first ozone level, second ozone level).

Two-way ANOVA with replication

Biology experiments are typically more complicated than a single t-test or one-way ANOVA design can handle; rarely would we conduct an experiment that reflects only one source of variation.

For example, while diet has a profound effect on weight, clearly, activity levels are also important. At a minimum, when considering a weight loss program, we would want to control or monitor activity of the subjects. This is a two-factor model, and the **main effects**, the two factors, were diet (factor A) and activity, (factor B). Both are expected to affect weight loss, and, perhaps, they may do so in complicated ways — an **interaction** (e.g., on DASH diet, weight loss is accelerated when subjects exercise regularly).

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + \epsilon_{ijk}$$

The subject of this chapter is the introduction to **two-way ANOVA** designs. In fact, to many, ANOVA design is practically synonymous to a statistician when they think about experimental design (Lindman 1992; Quinn and Keough 2002). As noted by Quinn and Keough (2002) in the preface to their book, "… many biological hypotheses, even deceptively simple ones, are matched by complex statistical models" (p. xv). Once you start adding factor variables there becomes a number of ways in which the groups and experimental units can be distributed, and thus impact the inferences one can make from the ANOVA results. The first statistical model we introduced was the one-way ANOVA. Next, we begin the two-way ANOVA with the **crossed, balanced, fully replicated design**. Along the way we introduce model symbols to help us communicate the design structure and implications of the statistical models.

14.1: Crossed, balanced, fully replicated designs

Introduction

“Biology is complicated” (p. 25, National Research Council [2005]), and as researchers we need to balance our need for statistical models that fit the data well and provide insight into the phenomenon in question against compressing that complexity in ways that do not reflect the phenomenon or hinder further progress in understanding the phenomenon. From our view as researchers then, we recognize that an experiment with only one causal variable is not likely to be informative. For example, while diet has a profound effect on weight, clearly, activity levels are also important. At a minimum, when considering a weight loss program, we would want to control or monitor activity of the subjects. This is a two-factor model; the two factors, diet and activity, are expected to both affect weight loss, and, perhaps, they may do so in complicated ways (e.g., on DASH diet, weight loss is accelerated when subjects exercise regularly).

Before we proceed, a word of caution is warranted. Prior to the 1990s, one could be excused for implementing experiments with simple designs that are suitable for analysis by contingency tables, t-tests, and one-way ANOVA. Now, with powerful computers available to most of us, and the feature-rich statistical packages installed on these computers, we can do much more complicated analyses on, hopefully, more realistic statistical models. This is surely progress, but caution is warranted nonetheless — just because you have powerful statistical tests available does not mean that you are free to use them — there is much to learn about the error structures of these more complicated models, for example, and how inferences are made across a model with multiple levels of **interaction**. In general it is preferred that experimental researchers consult and work with knowledgeable statisticians so that the most efficient and powerful experiment can be designed and subsequently analyzed with the correct statistical approach (Quinn and Keough 2002). Our introductory biostatistics textbook is not enough to provide you with all of the tools you would need, and while I do advocate self-learning when it comes to statistics, I do so provided we all agree that we are likely not getting the full picture this way. What we can do is provide an introduction to the field of experimental design with examples of classical designs so that the language and process of experimental design from a statistical point of view will become familiar and allow you to participate in the discussion with a statistician and read the literature as an informed consumer.

Two-factor ANOVA with replication

Our one factor statistical models can easily be extended to reflect more complicated models of causation, from one factor to two or more. We begin with two factors and the two-way ANOVA. Now we want to extend our discussion to examine how we can analyze data where we have two factors that may cause variation in the one response variable.

Consider the following two-way data set.

Table 14.1.1. Two-way data set of diet and population.

Diet A Population 1	Diet A Population 2	Diet B Population 1	Diet B Population 2
4	5	12	5
6	8	15	7
5	9	11	8

I’ve included the stacked version of this dataset at the end of this page (scroll to end or click here).

Question: What is the response variable? Which variable is the Factor variable? What are the classes of treatments and the levels of the treatments?

Answer.

Factors: Diet & Population

Levels: A, B for Diet;

Observations from population 1 or 2.

Note the replication: for every level of Diet (A or B) there is an equal number of individuals from the 2 populations. Said another way, there are three replicates from population 1 for Diet A, 3 replicated from population 2 for Diet A, etc.

And finally, we say that the experiment is CROSSED: Both levels of Diet have representatives of both levels of Population.

In order to properly analyze this type of research design (2 factor ANOVA, with equal replication), the data must be crossed. "Crossed" means that each level of Factor 1 must occur in each level of Factor 2.

From the example above: each population must have individuals given diet A and diet B.

Each of the collection of observations from the same combination of Factor 1 and Factor 2 is called a CELL:

All individuals in Diet A and Population 1 are in cell 1.

All individuals in Diet A and Population 2 are in cell 2.

All individuals in Diet B and Population 1 are in cell 3.

All individuals in Diet B and Population 2 are in cell 4.

If the data is completely crossed then you can calculate the number of cells:

Number of Levels in Factor 1 × Number of Levels in Factor 2 = Total Number of Cells

From the above example: 2 Diets × 2 Populations = 4 cells.

How to analyze two factors?

One solution (but inappropriate) is to do several separate One-Way ANOVAs.

There are two reasons that this approach is not ideal:

1. This approach will increase the number of tests performed and therefore will increase the chance of rejecting a Null Hypothesis when it is true (increase our p value without us being aware that it is changing – R and Rcmdr will not tell you there is a problem). This is analogous to the problems that we have seen if we perform multiple t-tests instead of a Mult-Sample ANOVA.
2. More importantly, there may be interactions among the TWO Factors in how they effect the response variable. One of the more interesting possible outcomes is that the influence of one of the Factors DEPENDS on the second FACTOR. In other words, there is an interaction between factor one and factor two on how the organism responds.

Here is a graph that illustrates one possible outcome:

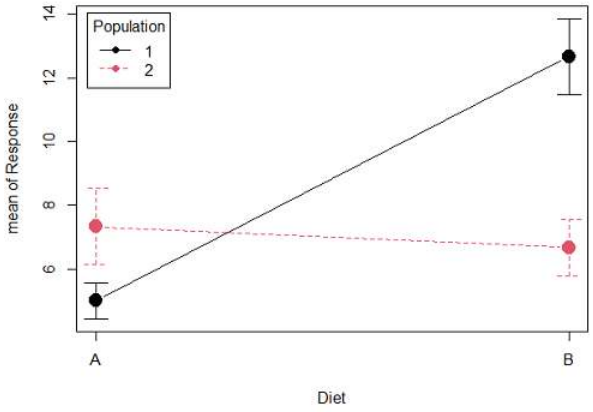


Figure 14.1.1*A*: One of several possible outcome of two treatments (factors). A clear interaction: First Diet level population 1 has greatest weight change, whereas for second diet level, population 2 has greatest weight change.

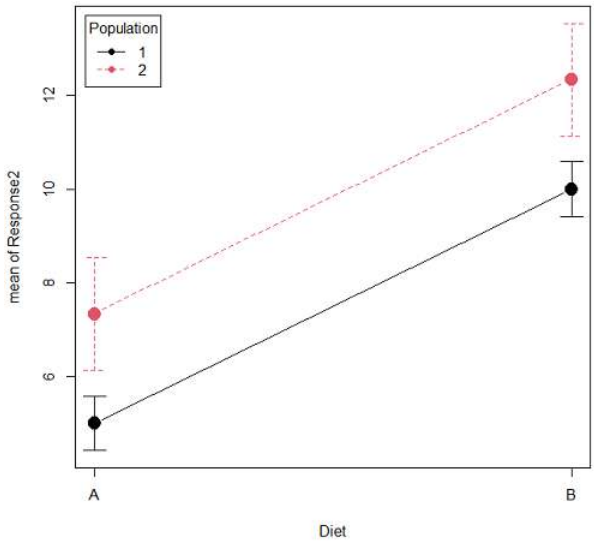


Figure 14.1.1*B*: One of several possible outcome of two treatments (factors). Clearly, no interaction: Population 1 always lower response than Population 2 regardless of Diet.

R code for plots:

Rcmdr: **Graphs → Plot of means…** then added `pch=19` and modified `legend.pos=` from `"farright"` to `"topleft"` .

Figure 14.1.1*A*:

```
with(pops2, plotMeans(Response, Diet, Population, pch=19, error.bars="se", connect=TRU
```

Figure 14.1.1*B*:

```
with(pops2, plotMeans(Response2, Diet, Population, pch=19, error.bars="se", connect=TR
```

Figure 14.1.1*A* and 14.1.1*B* shows that BOTH factors, Diet and Population, affect the Response of the subjects. Figure 14.1.1*A* also shows that the effects across Diet are not consistent: the responses are different. Individuals in Population 1 show decreased change in weight going from Diet A (1) to Diet B (2). But, individuals from Population 2 do just the opposite.

Figure 14.1.1*A*, because the effect of Diet cannot be interpreted without knowing which population you're looking at, shows what is called an **interaction** between Factor 1 and Factor 2. It's the part of the variation in the response NOT accounted for by either factor.

We can see the importance of doing the two-factor ANOVA by showing what would happen if we did two One-Factor (one-way) ANOVAs. For the first One-Factor (multi-sample) ANOVA we can examine the effect of Diet on weight. We could do this by combining the individuals from populations 1 & 2 that are given diet A (Diet A group) and then combining individuals from populations 1 & 2 that are given diet B (Diet B group).

An incorrect analysis of a two-way designed experiment

Statistical software will do exactly what you tell it to do, therefore, there is nothing to stop you from analyzing your two factor experimental design one variable at a time. It is statistical wrong to do so, but, again, there is nothing in the software that will prohibit this. So, we need to show you what happens when you ignore the experimental design in favor of a simple application of statistical analysis.

First, take a look at our two-way example with Diet as a factor and Population as another factor.

Here's is the one-way ANOVA for Diet only.

```
aov(Response ~ Diet, data=pops)
```

Table 14.1.2. One-way ANOVA table for diet (ignoring the other factor).

Source	DF	Sum of Squares	Mean Squares	F	P
Diet	1	36.75	36.75	4.26	0.066
Error	10	86.17	8.62		
Total	11	122.92			

When we ignore (combine) the identity of the two populations in this example, we see that it would APPEAR that Diet has NO EFFECT on the weight of the individuals, at least based on our statistical significance cut-off of Type I error set to 5%. Similarly, if we ignore Diet and compare responses by Population, the p-value was 0.367, not statistically significant (confirm p-value from one-way ANOVA on your own).

Now let's do the analysis correctly and pay attention to the main effect, Diet.

Here's the **2-way ANOVA table**.

```
lm(Response ~ Diet*Population, data=pops, contrasts=list(Diet ="contr.Sum", Populatio
+ ="contr.Sum"))
```

Table 14.1.3. Two-way ANOVA (the correct analysis!)

Source	DF	SS	MS	F	P
Diet	1	36.75	36.75	12.25	0.008
Population	1	10.08	10.08	3.36	0.104
Interaction	1	52.08	52.08	17.36	0.003
Error	8	24.00	3.00		
Total	11	122.92			

We can visualize the results by plotting the means for each treatment group (Fig. 14.1.2).

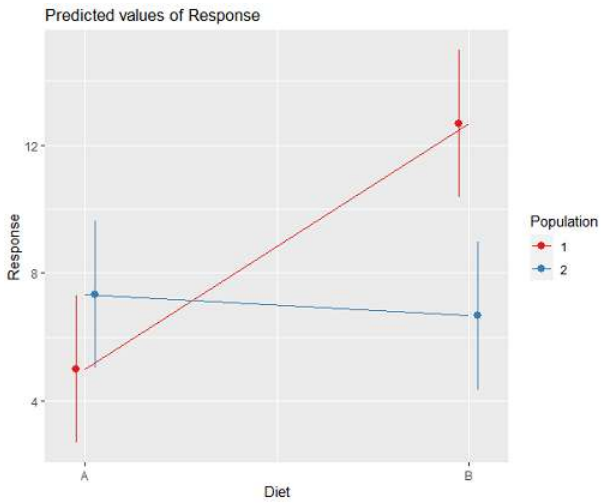


Figure 14.1.2: Plots of the main effects for Diet factor, levels A and B, and Population, levels 1 and 2.

R code for plot in Fig. 14.1.2:

```
library(sjPlot)
library(sjmisc)
library(ggplot2)
plot_model(LinearModel.1, type = "pred", terms = c("Diet", "Population")) + geom_line
```

And then for the interaction (Fig. 14.1.3).

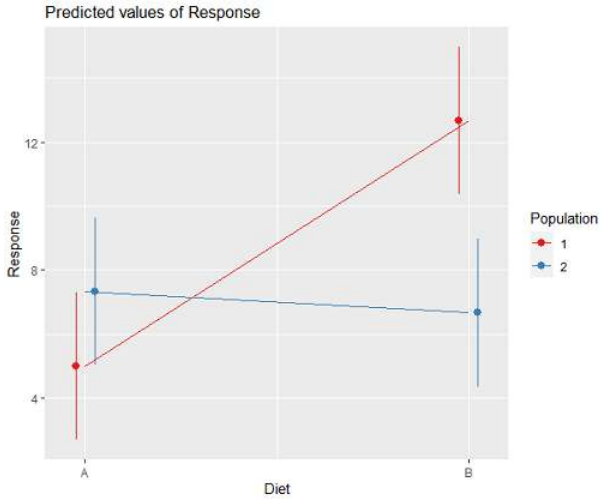


Figure 14.1.1: Copy and Paste Caption here. (Copyright; author via source)

R code: two-way ANOVA

The more general approach to running ANOVA in R is to use the general linear model function, `lm()`, saved as object `MyLinearModel.1`, for example, then follow up with

```
Anova(MyLinearModel.1, type="II")
```

to obtain the familiar ANOVA table. The `lm()` menu is obtained in `Rcmdr` by following **Statistics→ Fit models→ Linear model…**, and entering the model (Fig. 14.1.4). In this case, the model was $Response \approx Diet * Pop$

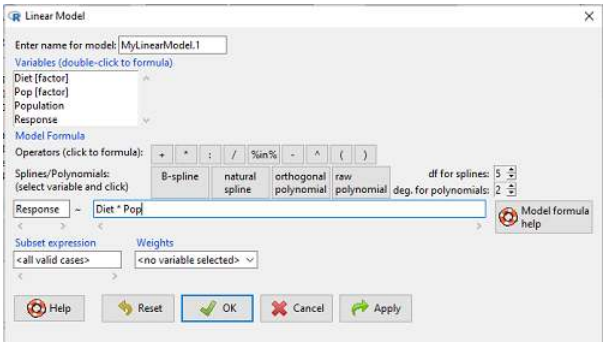


Figure 14.1.4: Linear model menu in Rcmdr.

```
Output from lm() function for this example

LinearModel.2 <- lm(Response ~ Diet * Pop, data=pops)
summary(LinearModel.2)

Call:
lm(formula = Response ~ Diet * Pop, data = pops)

Residuals:
Min 1Q Median 3Q Max
-2.3333 -1.1667 0.1667 1.0833 2.3333

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.000 1.000 5.000 0.00105 **
Diet[T.B] 7.667 1.414 5.421 0.00063 ***
Pop[T.2] 2.333 1.414 1.650 0.13757
Diet[T.B]:Pop[T.2] -8.333 2.000 -4.167 0.00314 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.732 on 8 degrees of freedom
Multiple R-squared: 0.8047, Adjusted R-squared: 0.7315
F-statistic: 10.99 on 3 and 8 DF, p-value: 0.003285
```

We want the ANOVA table, so run

```
Anova(MyLinearModel.1, type="II")
```

or in Rcmdr, **Models** → **Hypothesis tests** → **ANOVA table…** Accept the defaults (Types of tests = Type II, uncheck use of sandwich estimator), and press OK. I'll leave that for you to do (see Questions).

Interaction, explained

How can we **visualize** the effects of the Factors and the effects of the interaction? Plot the means of a two-factor ANOVA (Fig. 14.1.5). An interaction is present if the lines cross (even if they cross outside the range of the data), but if the lines are parallel, no interaction is present.

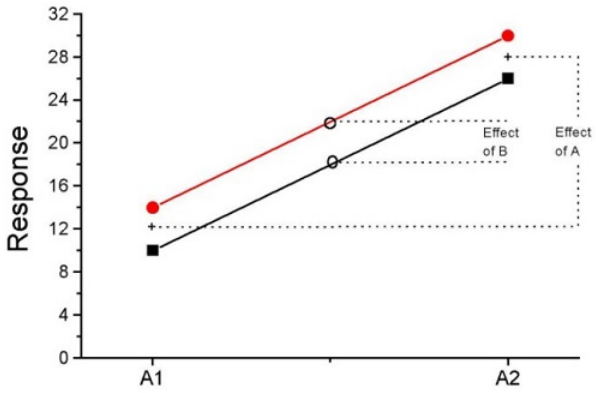


Figure 14.1.5: A plot showing no interaction between factor A and factor B for some ratio scale response variable.

A large effect of factor A – compare means

A small effect of factor B – compare means

Little or no interaction – lines are parallel

Three hypotheses for the Two-Factor ANOVA

The important advance in our statistical sophistication (from one to two factors!!) allows us to ask three questions instead of just two question:

1. Is there an effect of Factor 1?
   - H_O: There is no effect of Factor 1 on the response variable.
   - H_A: There is an effect of Factor 1 on the response variable.
2. Is there an effect of Factor 2?
   - H_O: There is no effect of Factor 2 on the response variable.
   - H_A: There is an effect of Factor 2 on the response variable.
3. Is there an INTERACTION between Factor 1 & Factor 2?
   - H_O: There is no interaction between Factor 1 & Factor 2 on the response variable.
   - H_A: There is an interaction between Factor 1 & Factor 2 on the response variable.

Questions

1. In the crossed, balanced two-way ANOVA, how many Treatment groups are there if Factor 1 has three levels and Factor 2 has four levels?
   A. 3
   B. 4
   C. 7
   D. 9
   E. 12
2. What is meant by the term “balanced” in a two-way ANOVA design?
   A. Within levels of a factor, each level has the same sample size
   B. Each level of one factor occurs in each level of the other factor
   C. There are no missing levels of a factor.
   D. Each level of a factor must have more than one sampling unit.
3. What is meant by the term “crossed” in a two-way ANOVA design?
   A. Within levels of a factor, each level has the same sample size
   B. Each level of one factor occurs in each level of the other factor
   C. There are no missing levels of a factor.
   D. Each level of a factor must have more than one sampling unit.
4. What is meant by the term “replicated” in a two-way ANOVA design?
   A. Within levels of a factor, each level has the same sample size

B. Each level of one factor occurs in each level of the other factor
C. There are no missing levels of a factor.
D. Each level of a factor must have more than one sampling unit.

5. Use the multi-way ANOVA command in `Rcmdr` to generate the ANOVA table for the example data set.
6. Use the linear model function and Hypothesis tests in `Rcmdr` to generate the ANOVA table for the example data set.

Data set

Don't forget to convert the numeric `Population` variable to character factor, e.g., a new object called `Pop` . The R command is simply

```
Pop <- as.factor(Population)
```

But easy to use `Rcmdr` also. From within `Rcmdr` select **Data → Manage variables in active dataset → Convert numeric variables to factors…**

Diet	Population	Response
A	1	4
A	1	6
A	1	5
A	2	5
A	2	8
A	2	9
B	1	12
B	1	15
B	1	11
B	2	5
B	2	7
B	2	8

14.2: Sources of variation

Introduction

Sources of variation, or components of the two-way ANOVA, include two factors, each with two or more levels (groups); collectively, factors are often referred to as the **main effects** in these types of ANOVA. The other source of variation in a two-way ANOVA is the **interaction** between the two factors. Below, I have listed the important components, although I have not included how the sum of squares are calculated. You are expected to know the sources of variation for this most basic two-way ANOVA table (Table 14.2.1). You should also be able to solve any missing elements in one of these tables by utilizing any included information.

Table 14.2.1. ANOVA table for two-way, balanced, replicated design.

Source of variation	Degrees of Freedom	Mean Squares	F-statistic
Factor A (Diet)	$a-1$	$\frac{factor\ A\ SS}{factor\ A\ DF}$	$\frac{factor\ A\ MS}{error MS}$
Factor B (Population)	$b-1$	$\frac{factor\ B\ SS}{factor\ B\ DF}$	$\frac{factor\ B\ MS}{error MS}$
$A \times B$ interaction	$(a-1)(b-1)$	$\frac{A \times B\ SS}{A \times B\ DF}$	$\frac{A \times B\ MS}{error MS}$
Error (within-cells)	$ab(n-1)$	$\frac{error\ SS}{error\ DF}$	
Total	$N-1$		

Taking each row from Table 14.2.1 one at a time, we have:

Source of variation	Degrees of Freedom	Mean Squares	F-statistic
First Factor	$a-1$	$\frac{factor\ A\ SS}{factor\ A\ DF}$	$\frac{factor\ A\ MS}{error MS}$

where Source refers to the source of variation, DF refers to Degrees of Freedom, a is the number of levels (groups) of the first factor, SS refers to Sum of Squares, and MS refers to the Mean Squares.

Next is the second factor

Source of variation	Degrees of Freedom	Mean Squares	F-statistic
Second Factor	$b-1$	$\frac{factor\ B\ SS}{factor\ B\ DF}$	$\frac{factor\ B\ MS}{error MS}$

where b is the number of levels (groups) of the second factor. Next is the interaction between the first and second factors.

Source of variation	Degrees of Freedom	Mean Squares	F-statistic
Interaction	$(a-1)(b-1)$	$\frac{A \times B\ SS}{A \times B\ DF}$	$\frac{A \times B\ MS}{error MS}$

and lastly the Within-cell Error or residual source of variation

Source of variation	Degrees of Freedom	Mean Squares	F-statistic
Error (within-cells)	$ab(n-1)$	$\frac{error\ SS}{error\ DF}$	N/A

where n is the number of experimental units for each group. Note that if the sample size differs for one or more groups (levels),then the design would be unbalanced and this formula does not work to determine the degrees of freedom. The total degrees of freedom for the two-way ANOVA is simply $N-1$, where N is the sample size for the entire problem; a little algebra shows that N may be calculated as $N = abn$

Unbalanced designs

An **unbalanced design** implies that observations are missing value for one or more groups. What to do if data are missing? The decision depends on how the data are missing (see Chapter 5). For example, if data are missing at random with respect to treatment, then this should not affect inference. If data are missing not at random, then inference, logically, must be impacted. Calculating the ANOVA, moreover, becomes a different matter. In the one-way ANOVA, no real problem arises although setting up contrasts among the levels requires a weighting term to be factored into the calculations. For higher-level ANOVA involving two or more factors, the sums of squares for treatment effects are no longer simple partitioning into the different sources of variation. The sources overlap and the order by which the Factors enter into the statistical model now affects the calculations. Thus, while setting up the calculations for the balanced design is straightforward, perhaps surprisingly, if group sizes differ, this simple relationship for calculating the degrees of freedom, sums of squares, and Mean squares becomes an unsolvable problem. This problem is largely solved by the **general linear model**.

Questions

1. In two-way ANOVA, what should you always test first?
   A. The significance of Factor 1.
   B. The significance of Factor 2.
   C. The significance of the interaction between Factor 1 and Factor 2.
   D. Doesn't matter which is tested first because you have three null hypotheses in the 2-way ANOVA.
2. Why is the cell empty for F statistic in the Within-cell Error or residual source of variation?
3. Based on the results of a two-way ANOVA, the error sums of squares (SSE) was computed to be 160. If we ignore one of the factors and perform a one-way ANOVA using the same data, will the SSE be the same as in the two-way ANOVA, or will it increase? Decrease? Explain your choice.
4. While conducting a two-way ANOVA, you conclude that a statistically significant interaction exists between factor 1 and factor 2. What should be your next step? Do you drop the interaction term from the model and redo the analysis or do you report the results of factor effects including the non-significant interaction?

14.3: Fixed effects, random effects

Introduction

With few exceptions (e.g., repeatability and intraclass correlation calculations, Chapter 12.3), we have been discussing the Model I ANOVA or **fixed-effects** ANOVA — fixed implies that we select the levels for the factor. It may not be obvious — in hindsight it is — but levels may also be randomly selected, e.g., nature provides the levels. Thus, levels are random and the model is a **random-effects** ANOVA. Beginning with our discussions of ANOVA, it becomes increasingly important to incorporate concept of models in statistics. As you have been working in R and Rcmdr with the `lm()` function, you have been forced to address the statistical model concept — you enter the response variable then type in both factors and create a term for the interaction.

We've just completed an experiment in which the response (cholesterol levels) of 12 individuals from one of two drug treatments (1 or 2), given one of 2 types of diets (Diet A or Diet B), was observed. Thus, we say that the specific treatments of drug and diet contribute to variation in cholesterol levels. More formally, we say that the observed response of the k^{th} individual (Y_{ijk}) is equal to the overall mean (μ) plus the added effect of Drug (α) plus the added effect of Diet (β) plus the interaction between Diet and Drug Population $(\alpha\beta)$ plus unidentified sources of variation generally called error (ϵ). In symbols, we write

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where i is the number of levels of the first factor (in our example, Population had 2 levels, so $i = 1$ or $i = 2$), j is the number of levels of the second factor (in our example, Diet had 2 levels, so $j = A$, or $j = B$), and k is the total number of observations in the experiment (12 in our example, so $k = 1, 2, \ldots, 11, 12$). Thus, we can think of each term "adding up" to give us the observed value.

Although it is a bit confusing at first, these equations help us understand how the experiment was conducted and therefore how to analyze and interpret the results.

Model I, Model II & Model III ANOVA

Statisticians recognize that how levels of the factors were selected for an experiment impact conclusions from ANOVAs. The key: whether or not the levels of the factor were selected (1) randomly from all possible levels of the factor or (2) specifically selected by the experimenters. We introduced the concepts of fixed and random effects in Chapter 12.3. For one-way ANOVA, the distinction between fixed and random effects influences the interpretation, but not the calculation, of the ANOVA components. For two or more treatment factors, both the interpretations and the calculations of ANOVA components are affected.

There are two general types of Factors that we can choose to employ in an ANOVA: Fixed Model ANOVA and Random Model ANOVA. Where two or more factors apply, by far the most common model in experimental sciences is a combination of fixed and random, so we need to add a third general type, the Mixed Model ANOVA design.

Fixed Factors. Where the levels of the factor are selected by us. In this case we would only be interested in the response of the individuals that are given those specific treatments.

> **Medicine** – for example, where we choose a treatment given to patients with a history of coronary heart disease; compare outcomes of patients given a statin (drug used to lower serum cholesterol) drug versus a placebo. (Note that this is the same study we discussed in our lecture on about risk analysis).
>
> **Ecotoxicology** – for example, compare growth rates and deformities of tadpole frogs given Aldicarb, Atrazine (both estrogen-mimicking pesticides), or a control (i.e., no pesticides). If you're interested in these topics, here's a link to the EPA's web site, listing pesticide sales and use in the United States. Here's a link to a NIH National Institute of Environmental Sciences, with a nice description of estrogen mimicking pesticides.
>
> **Agriculture & Genetics** – for example, monitor growth of a particular hybrid corn available from three different manufacturers. See an example of such studies here.

In each of these examples we might be interested in those specific treatments and no other treatments.

> H_O: No difference in the means among the levels of the Factor
>
> H_A: Some difference in the means among these specific levels of the Factor; the specific levels of this factor effect the response variable.

This is an example of a **Model I ANOVA**, also called a "**fixed effects**" model ANOVA.

Random Factors. We still only use a relatively few number of different levels of a particular factor. However, in this case we are interested in many different levels of the factor — we want to generalize beyond our sample. The levels that we use would be a

"sample" of all possible levels that we would be interested in.

Medicine – for example, where we randomly choose a treatment level given to patients; four concentrations of a drug and a placebo. Since concentration is a ratio scale data type, concentration can range from 0 (the placebo) to 100%.

Ecotoxicology – for example, release different concentrations or mixtures of air plus components of air pollution to chambers, record the response of plants or animals.

Agriculture & Genetics – for example, grow three different varieties of a plant in three different soils or different genetic strains of animals on three different diets. In these cases, factor levels are random because we are drawing from a large pool of possible levels: genetic varieties or strains — we selected three, but it's rare that were are specifically interested in the three chosen. More often, we want to make generalizations and the three were somehow representative (we hope) of genetic variation in the species of interest.

In each of these examples, we write the null hypothesis to reflect that the particular levels are only of interest in so far as they can be used to generalize back to the population.

H_O: No difference in the variation between groups.

H_A: Some difference in the variation among these groups.

This is an example of a **Model II** ANOVA, also called a "**random effects**" model ANOVA. Note that we specify the hypothesis in terms of variation, not of the means.

Your two-way ANOVA could be Model I, Model II, or it could be **mixed**, with one factor fixed, the other random (this later model is called a **Model III**, or "**mixed model**" ANOVA).

For the most part, the distinction between whether you have fixed or random effects is clear, but whether we use fixed or random or combinations, **this design decision has consequences for testing**.

The decision does not affect the **Sources of Variation** for the different Models. Last time, we showed the tests for a Model I ANOVA (the "fixed effects model").

For Random Effects or Mixed Effects, **we only change how** we determine the statistical significance of the Factors. Here's a summary of how the experimental design changes the calculation of the F-test statistic for the Factors.

Table 14.3.1. Calculation of F for different experimental designs.

Factor	Model I: Both factors fixed	Model II: Both factors random	Model III: Factor A fixed, Factor B random	Model III: Factor A random, Factor B fixed
A	$\frac{factor\ A\ MS}{error\ MS}$	$\frac{factor\ A\ MS}{A \times B\ MS}$	$\frac{factor\ A\ MS}{A \times B\ MS}$	$\frac{factor\ A\ MS}{error\ MS}$
B	$\frac{factor\ B\ MS}{error\ MS}$	$\frac{factor\ B\ MS}{A \times B\ MS}$	$\frac{factor\ B\ MS}{error\ MS}$	$dfracfactor\ B\ MSA \times B\ 1$
A × B	$\frac{A \times B\ MS}{error\ MS}$	$\frac{A \times B\ MS}{error\ MS}$	\(\frac{A \times B \ MS}{error \ MS}\	\(\frac{A \times B \ MS}{error \ MS}\

The Critical Value for each of the different F values will be obtained by simply finding the degrees of freedom for the numerator and denominator SS. This was discussed and can be found in the section on sources of variation in 2-way ANOVA.

From the formulas, we can see that the major difference is that sometimes the Factor MS is divided by the error MS and sometimes it is divided by the interaction MS.

If the interaction term is NOT statistically significant, then the Interaction MS (mean square) estimates the Error MS. In other words, if the interaction term is not statistically significant it will be similar in magnitude to the Error MS. In this case there will be no large difference in the computed outcomes if the Factor A or B is fixed or random.

However, there will be times when the interaction is not significant but the interaction MS is still larger than the Error MS. Then there could be a difference in the F value for the Factor.

If the interaction term is Significant and the interaction MS is larger than the Error MS then there will be difference in F values for the Factors A and/or B. The F values will be smaller for the Factors MS that are divided by the interaction MS. It is possible that they will become non-significant with the interaction MS as the denominator. Therefore, it will become harder to detect a significant Factor

effect if there is also a significant Interaction effect. A graphical representation will help us understand why we use the interaction MS in some instances as the denominator.

In fact, if the interaction is found to be statistically significant, we must then interpret the effects of factors with caution. In general, if the interaction is significant, then the main factors are generally not interpreted in the 2-way ANOVA. Instead, a series of one-way ANOVAs are conducted holding one of the factors constant. For example, evaporative water loss (EWL) in frogs in the presence of air pollution (ozone) may depend on the relative humidity (RH) — if the RH is low, the frog may lose less body water at different concentrations of ozone than if the RH is moderate. Therefore, since the interaction is significant, the best thing to do is to look at the effects of ozone concentration on EWL at each level of saturation (RH).

This is a critical point in your understanding of complex ANOVA designs. Let us examine a case where there is a mildly significant interaction effect between two factors. In the first graph below (Fig. 14.3.1) we see that Genotype 2 performs better on average (combining the two density treatments). If we are only interested in these two density treatments then it might be that the Genotype Factor is significant. This would be the case if Factor A is fixed and Factor B (density) is also fixed.

The Formula for Factor A would be: $F = Genotype\ MS / error\ MS$

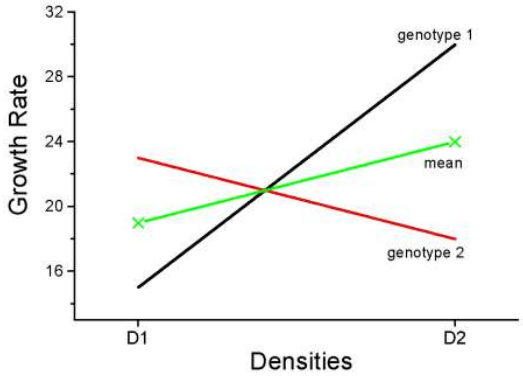


Figure 14.3.1: Interaction example. At density D1, genotype 2 (red line) has higher growth rate; at density D2, the ranking switches: now, genotype 1 (**black line**) has higher growth rate.

However, it is likely that both Factor A (genotype) and Factor B (densities) are actually "samples" of many other possible genotypes and densities that we could examine.

Consider more than two genotypes raised in more than 2 densities (Fig. 2). The outcome might look like

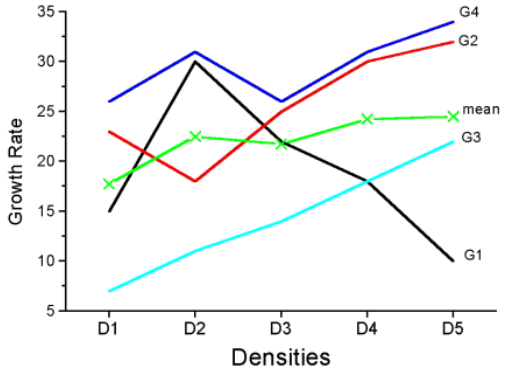


Figure 14.3.2: Interaction example expanded for multiple genotypes over multiple densities.

The graph (Fig. 14.3.2) shows that genotype 2 does not do better than genotype 1 if we have more densities. If we also have other genotypes we see that there are other genotypes that have better (higher) responses than genotype 2.

$$F = \frac{Genotype\ MS}{Interaction\ MS}$$

In these cases it would have been more appropriate to calculate the F value for Factor A (genotype) using the interaction MS as the denominator. In Figure 14.3.1 there was some interaction this will make it harder to reject the null hypothesis that there is no effect of Factor A (genotype). So you must be careful to think about how you plan to interpret your data before you decide how to analyze the data using a Two-Factor ANOVA.

Questions

1. Which of the following statements regarding fixed and random factors is true?
   A. With fixed factors, the subjects are selected by the researcher
   B. With fixed factors, the treatment levels are selected by the researcher at random from all possible levels
   C. With fixed factors, the subjects are selected at random by the researcher
   D. With random factors, the treatment levels are selected by the researcher at random from all possible levels
2. Please write the equation for the one-way ANOVA with four levels of of fixed effects treatment factor A (you may wish to review Chapter 12.2)
3. Selecting from all possible levels of a statin drug would be an impossible and meaningless experimental design. Explain why.
4. For a multiway ANOVA design, when will the differences in the Random versus Fixed Factor make a difference?

14.4: Randomized block design

Introduction

Randomized Block Designs and Two-Factor ANOVA

In previous lectures, we have introduced you to the standard factorial ANOVA, which may be characterized as being **crossed**, **balanced**, and **replicated**. We expect that additional factors (**covariates**) may contribute to differences among our experimental units, but rather than testing them — which would increase the need for additional experimental units because of the increased number of groups to test — we randomize our subjects. Randomization is intended to disrupt trends of **confounding variables** (aka covariates). If the resulting experiment has **missing values** (see Chapter 5), then we can say that the design is partially replicated; if only one observation is made per group, then the design is not replicated — and perhaps, not very useful!!

A special type of Two-factor ANOVA which includes a "blocking" factor and a treatment factor.

Randomization is one way to control for "uninteresting" confounding factors. Clearly, there will be scenarios in which randomization is impossible. For example, it is impossible to randomly assign subjects to The **blocking factor** is similar to the Paired t-test. In the paired t-test we had two individuals or groups that we paired (e.g. twins). One specific design is called the Randomized Block Design and we can have more than 2 members in the group. We arrange the experimental units into similar groups, i.e., by the blocking factor. Examples of blocking factors may include day (experiments may be run over different days), location (experiments may be run at different locations in the laboratory), nucleic acid kits (different vendors), operator (different assistants may work on the experiments), et cetera.

In general we may not be directly interested in the blocking factor. This blocking factor is used to control some factor(s) that we suspect might affect the response variable. Importantly, this has the effect of reducing the sums of squares by an amount equal to the sums of squares for the block. If variability removed by the blocking is significant, Mean Square Error (MSE) will be smaller, meaning that the F value for treatment will be bigger — meaning we have a more powerful ANOVA than if we had ignored the blocking.

Statistical Testing in Randomized Block Designs

"Blocks" is a **Random Factor** because we are "sampling" a few blocks out of a larger possible number of blocks. Treatment is a Fixed Factor, usually.

The statistical model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon i, j$$

The Sources of Variation are simpler than the more typical Two-Factor ANOVA because we do not calculate all the sources of variation – the interaction is not tested! (Table 14.4.1).

Table 14.4.1. Sources of variation for a two-way ANOVA, randomized block design.

Sources of Variation & Sum of Squares	DF	Mean Squares
$total\ SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(X_{ijk} - \bar{X}\right)^2$	$N-1$	$\frac{total\ SS}{total\ DF}$
$SS_{Factor\ A} = b \cdot n \sum_{i=1}^{a}\left(\bar{X}_i - \mu\right)^2$	$a-1$	$\frac{SS\ Factor\ A}{DF\ Factor A}$
$SS_{Factor\ B} = a \cdot n \sum_{j=1}^{b}\left(\bar{X}_j - \mu\right)^2$	$b-1$	$\frac{SS\ Factor\ B}{DF\ Factor B}$
$SS_{Total} = SS_{Factor\ A} - SS_{Factor\ B}$	$n-a-b$	$\frac{SS\ Remainder}{DF\ Remainder}$

Critical Value $F_{0.05(2),(a-1),(Total\ DF-a-b)}$

In the exercise example above: Factor A = exercise or management plan.

Notice that we do not look at the interaction MS or the Blocking Factor (typically).

Learn by doing

Rather than me telling you, try on your own. We'll begin with a worked example, then proceed to introduce you to three problems. See Chapter 14.8 for general discussion of `Rcmdr` and linear models for models other than the standard 2-way ANOVA.

Worked example

Wheel running by mice is a standard measure of activity behavior. Even wild mice will use wheels (Meijer and Roberts 2014). For example, we conduct a study of family parity to see if offspring from the first, second, or third sets of births perform differently in wheel-running behavior (measured in total revs per 24 hr period). Each set of offspring from a female could be treated as a block. Data are for 3 female offspring from each pairing. This type of data set would look like this:

Table 14.4.2. Wheel running behavior (revolutions of wheel per 24-hr period) by three offspring from each of three birth cohorts among four maternal sets (moms).

Block	Dam 1	Dam 2	Dam 3	Dam 4
b1	1100	1566	945	450
b1	1245	1478	877	501
b1	1115	1502	892	394
b2	999	451	644	605
b2	899	405	650	612
b2	745	344	605	700
b3	1245	702	1712	790
b3	1300	612	1745	850
b3	1750	508	1680	910

Thus, there were nine offspring for each female mouse (Dam1 – Dam4), three offspring per each of three litters of pups. The litters are the blocks. We need to get the data stacked to run in R. I've provided the dataset for you, so scroll to end of this page or click here.

Question 1. Describe the problem and identify the treatment and blocking factors.

Answer. Each female has three litters. We're primarily interested in genetics (and maternal environment) of wheel running behavior, which is associated with the moms (Treatment factor). The questions is whether there is an effect of birth parity on wheel running behavior. Offspring of a first-time mother may experience different environment than offspring of an experienced mother. In this case, parity effects is an interesting question; nevertheless, blocking is the appropriate way to handle this type of problem.

Question 2. What is the statistical model?

Answer. Response variable, Y, is wheel running. Let α be the effect of Dam and β the birth cohorts (i.e., the blocking effect).

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{i,j}$$

Question 3. Test the model.

Answer. We fit the main effects (Dam and Block). $Wheel \sim Dam + Block$

Rcmdr: Statistics → Fit models → Linear model

Figure 14.4.1: Enter formula for the linear model in R Commander.

then run the ANOVA summary to get the ANOVA table. **Rcmdr: Models → Hypothesis tests → ANOVA table.**

Output

```
Anova Table (Type II tests)

Response: Wheel
           Sum Sq  Df  F value    Pr(>F)
Dam       1467020   3   4.4732   0.01036 *
Block     1672166   2   7.6482   0.00207 **
Residuals 3279544  30
```

Question 4. Conclusions?

Answer. The null hypotheses are:

Treatment factor: Offspring of the different dams have same wheel running activity of offspring.

Blocking factor: No effect of litter parity on wheel running activity of offspring.

Both the treatment factor (p = 0.01036) and the blocking factor (p = 0.00207) were statistically significant.

Problem 1.

Or we might want to measure the Systolic Blood Pressure of individuals that are on different exercise regimens. However, we are not able to measure all the individuals on the same day at the same time. We suspect that time of day and the day of the year might affect an individual's blood pressure. Given this constraint, the best research design in this circumstance is to measure one individual on each exercise regime at the same time. These different individuals will then be in the same "block" because they share in common the time that their blood pressure was measured. This type of data set would look like this (Table 14.4.2):

Table 14.4.2. Simulated blood pressure of five subjects on three different exercise regimens.†

Subject	No Exercise	Moderate Exercise	Intense Exercise
1	120	115	114
2	135	130	131
3	115	110	109
4	112	107	106
5	108	103	102

†You'll need to arrange the data like the data set for the worked example.

Question 1. Describe the problem and identify the treatment and blocking factors.

Question 2. What is the statistical model?

Question 3. Test the model.

Question 4. Conclusions?

Problem 2.

Another example in conservation biology or agriculture. There may be three different management strategies for promoting the recovery of a plant species. A good research design would be to choose many plots of land (blocks) and perform each treatment (management strategy) on a portion of each plot of land (block). A researcher would start with an equal number of plantings in the plots and see how many grew. The plots of land (blocks) share in common many other aspects of that particular plot of land that may affect the recovery of a species.

Table 14.4.3. Growth of plants in 5 different plots subjected to one of three management plans (simulated data set).†

Plot No.	No Management Used	Management Plan 1	Management Plan 2
1	0	11	14

Plot No.	No Management Used	Management Plan 1	Management Plan 2
1	3	11	19
2	2	13	15
3	3	11	19
4	4	10	16
5	5	15	12

†You'll need to arrange the data like the data set for the worked example.

Plot No. | No Management Used | Management Plan 1 | Management Plan 2
1 | 3 | 11 | 19
2 | 2 | 13 | 15
3 | 3 | 11 | 19
4 | 4 | 10 | 16
5 | 5 | 15 | 12
†You'll need to arrange the data like the data set for the worked example.

These are examples of Two-Factor ANOVA but we are usually only interested in the treatment Factor. We recognize that the blocking factor may contribute to differences among groups and so wish to control for the fact that we carried out the experiments at different times (e.g., seasons) or at different locations (e.g., agriculture plots)

Question 1. Describe the problem and identify the treatment and blocking factors.

Question 2. What is the statistical model?

Question 3. Test the model.

Question 4. Conclusions?

Repeated-Measures Experimental Design

If multiple measures are taken on the same individual, then we have a repeated-measures experiment. This is a Randomized Block Design. In other words, each animal gets all levels of a treatment (assigned randomly). Thus, samples (individuals) are not independent and the analysis needs to take this into account. Just like for paired *t*-tests, one can imagine a number of experiments in biomedicine that would conform to this design.

Problem 3.

The data are total blood cholesterol levels for 7 individuals given 3 different drugs (from example, as given in Zar 1999, Ex 12.5, pp. 257-258).

Table 14.4.4. Repeated measures of blood cholesterol levels of seven subjects on three different drug regimens.†

Subjects	Drug 1	Drug 2	Drug 3
1	164	152	178
2	202	181	222
3	143	136	132
4	210	194	216
5	228	219	245
6	173	159	182
7	161	157	165

†You'll need to arrange the data like the data set for the worked example.

Question 1: Is there an interaction term in this design?

Question 2: Are individuals a fixed or a random effect?

Question 2. What is the statistical model?

Question 3. Test the model. Note that we could have done the experiment with 21 randomly selected subjects and a one-factor ANOVA. However, the repeated measures design is best IF there is some association ("correlation") between the data in each row. The computations are identical to the randomized block analysis.

Question 4. Conclusions?

Problem 4.

Here is a second example of a repeated measures design experiment. Garter snakes respond to odor cues to find prey. Snakes use their tongues to "taste" the air for chemicals, and flick their tongues rapidly when in contact with suitable prey items, less frequently for items not suitable for prey. In the laboratory, researchers can test how individual snakes respond to different chemical cues by presenting each snake with a swab containing a particular chemical. The researcher then counts how many times the snake flicks its tongue in a certain time period (data presented p. 301, Glover and Mitchell 2016).

Table 14.4.5. Tongue flick counts of naïve newborn snakes to extracts†

Snake	Control (dH_2O)	Fish mucus	Worm mucus
1	3	6	6
2	0	22	22
3	0	12	12
4	5	24	24
5	1	16	16
6	2	16	16

†You'll need to arrange the data like the data set for the worked example.

Question 1. Describe the problem and identify the treatment and blocking factors.

Question 2. What is the statistical model?

Question 3. Test the model.

Question 4. Conclusions?

Additional questions

1. The advantage of a randomized block design over a completely randomized design is that we may compare treatments by using _____ experimental units.
   A. randomly selected
   B. the same or nearly the same
   C. independent
   D. dependent
   E. All of the above
2. Which of the following is NOT found in an ANOVA table for a randomized block design?
   A. Sum of squares due to interaction
   B. Sum of squares due to factor 1
   C. Sum of squares due to factor 2
   D. None of the above are correct
3. A clinician wishes to compare the effectiveness of three competing brands of blood pressure medication. She takes a random sample of 60 people with high blood pressure and randomly assigns 20 of these 60 people to each of the three brands of blood pressure medication. She then measures the decrease in blood pressure that each person experiences. This is an example of (select all that apply)
   A. a completely randomized experimental design
   B. a randomized block design
   C. a two-factor factorial experiment
   D. a random effects or Type II ANOVA
   E. a mixed model or Type III ANOVA
   F. a fixed effects model or Type I ANOVA
4. A clinician wishes to compare the effectiveness of three competing brands of blood pressure medication. She takes a random sample of 60 people with high blood pressure and randomly assigns 20 of these 60 people to each of the three brands of blood pressure medication. She then measures the blood pressure before treatment and again 6 weeks after treatment for each person. This is an example of (select all that apply)
   A. a completely randomized experimental design
   B. a randomized block design
   C. a two-factor factorial experiment
   D. a random effects or Type II ANOVA
   E. a mixed model or Type III ANOVA
   F. a fixed effects model or Type I ANOVA

Data sets used in this page

Worked Example data set

Block	Dam	Wheel
B1	D1	1100
B1	D2	1566
B1	D3	945
B1	D4	450
B1	D1	1245
B1	D2	1478
B1	D3	877
B1	D4	501
B1	D1	1115
B1	D2	1502
B1	D3	892
B1	D4	394
B2	D1	999
B2	D2	451
B2	D3	644
B2	D4	605
B2	D1	899
B2	D2	405
B2	D3	650
B2	D4	612

Block	Dam	Wheel
B2	D1	745
B2	D2	344
B2	D3	605
B2	D4	700
B3	D1	1245
B3	D2	702
B3	D3	1712
B3	D4	790
B3	D1	1300
B3	D2	612
B3	D3	1745
B3	D4	850
B3	D1	1750
B3	D2	508
B3	D3	1680
B3	D4	910

14.5: Nested designs

Introduction

Crossed versus nested design

Factors are independent variables whose values we control and wish to study because we believe they have an effect on the dependent variable. While it is logical to think of factors and **levels** within factors as independent variables fully under our control, a moments reflection will come up with examples in which the groups (levels) depend on the factor.

Crossed – each level of a factor is in each level of the other factor. This was illustrated in Chapter 14.1 on the crossed, balanced, fully replicated two-way ANOVA.

Nested – levels of one factor are NOT the same in each of the levels of the other factor. Nested designs are an important experimental design in science, and they have some advantages over the 2-way ANOVA design (for one), but they also have limitations.

Classic examples of nesting: culturing and passage of cell lines in routine cell colony maintenance means that even repeated experiments are done on different experimental units. Cells derived from one vial are different from cells derived from a different vial. Similarly, although mice from an inbred strain are thought to be genetically identical, environments vary across time, so mice from the same strain but born or purchased at different times are necessarily different. These scenarios involving time create a natural **block effect**. Thus, cells are nested by block effect passage number and mice are nested by block effect colony time. We introduced randomized block design in the previous section, Chapter 14.4.

Statistical model

If Factor B is nested within Factor A, then a group or level within Factor B occurs only within a level of Factor A. Like the randomized block model, there will be no way to estimate the interaction in a nested two-way ANOVA. Our statistical model then is

$$Y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk}$$

Examples

Example 1. Three different drugs, 6 different sources of the drugs. The researcher obtains three different drugs from 6 different companies and wants to know if one of the drugs is better than another drug (Factor A) in lowering the blood cholesterol in women. There is always the possibility that different companies will be better or worse at making the drug. So the researchers also use the Factor Source (Factor B) to examine this possibility. Unfortunately they can not obtain all drugs from the same sources. This leads to a **Nested ANOVA** — notice that each drug is obtained from a different source.

We CANNOT perform the typical **two-factor ANOVA** because we cannot get a mean of the different drugs by combining the same levels of the Sources: the data is NOT crossed. The Sources of the drugs (Factor B) are NESTED within the type of Drug (Factor A): each source is only found in one of the Drug categories. So, we can't calculate a mean for the Drug levels independent of the SOURCE from which the drug came.

Table 14.5.1. Example of a nested design.

Drug A		Drug B		Drug C	
Source 1	Source 2	Source 3	Source 4	Source 5	Source 6
202.6	189.3	212.3	203.6	189.1	194.7
207.8	198.5	204.4	209.8	219.9	192.8
190.2	208.4	221.6	204.1	196.0	226.5
211.7	205.3	209.2	201.8	205.3	200.9
201.5	210.0	222.1	202.6	204.0	219.7

Scroll to end of this page to get the data set in **stacked worksheet** format, or click here.

Compare Table 14.5.1 to CROSSED data structure (Table 14.5.2) — a typical two-factor ANOVA — which would look like

Table 14.5.2. Contents of Table 14.5.1 presented as crossed design.

Drug A		Drug B		Drug C	
Source 1	Source 2	Source 1	Source 2	Source 1	Source 2
202.6	189.3	?	?	?	?
207.8	198.5	?	?	?	?
190.2	208.4	?	?	?	?
211.7	205.3	?	?	?	?
201.5	210.0	?	?	?	?

We can take a mean of the different drugs by combining the same levels of the Sources. Here's the nested design (Table 14.5.3).

Table 14.5.3. Group means, nested design.

Drug A		Drug B		Drug C	
Source 1	Source 2	Source 3	Source 4	Source 5	Source 6
202.76	202.3	213.92	204.38	202.86	206.92

We can take a mean of the different drugs by combining the same levels of the Sources. Here's the crossed design (Table 14.5.4).

Drug A		Drug B		Drug C	
Source 1	Source 2	Source 1	Source 2	Source 1	Source 2
202.76	202.3	?	?	?	?

Why the "?" in Tables 14.5.2 and 14.5.4? Manufacturing source 1 & 2 do not sell Drug B and Drug C. So, there cannot be a crossed design.

Why can't we just use a One-Way ANOVA? Can't we just ANALYZE the three DRUGS separately, ignoring the source issue (after all, the drugs are not all made by the same manufacturer)? But it is not a one-way ANOVA problem… Here's why.

The researcher suspects that the response of a particular drug might be dependent upon the particular source from which the drug was purchased. So, the type of source from which the drug was purchased is another FACTOR. Thus, drugs from one source might have more (less) affect compared to drugs from another source regardless of the type of drug. However, each drug is NOT available from each source. Thus the research design can NOT be crossed and Drug is NESTED within Source.

We can ask ONLY two questions (hypotheses) from this NESTED ANOVA research design:

H_O: There is no difference in the average effect of the drugs on (tumor size, cholesterol level, blood pressure, etc.)

H_A: There is a difference in the average effect of the drugs on (tumor size, cholesterol level, blood pressure, etc.)

H_O: There is no difference in the average effect of the drugs on (tumor size, cholesterol level, blood pressure, etc.) purchased from different manufacturers.

H_A: There is a difference in the average effect of the drugs on (tumor size, cholesterol level, blood pressure, etc.) purchased from different manufacturers.

Notice that we do NOT examine the effect of the interaction between Drug type and source of the drug. Why not?

Table $PageIndex5$. Sources of Variation in Nested ANOVA.

Source of Variation	Sum of Squares	DF	Mean Squares
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk}-\mu\right)^2$	$N-1$	

Source of Variation	Sum of Squares	DF	Mean Squares
Among all subgroups	$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij}-\mu\right)^2$	$ab-1$	
Among groups	$\sum_{i=1}^{a}\left(X_i-\mu\right)^2$	$a-1$	$\frac{among\ groups\ SS}{among\ groups\ DF}$
Among subgroups	$\sum_{i=1}^{a}\sum_{j=1}^{b}n_{ij}(X_i-\mu)^2$	$a(b-1)$	$\frac{among\ subgroups\ SS}{among\ subgroups\ DF}$
Error	Subtract all of the subgroup Sums of Squares from the Total Sums of Squares	$N-ab$	$\frac{error\ SS}{error\ DF}$

Testing nested ANOVA with one main factor

Perhaps surprisingly given the number of terms above, there are only two hypothesis tests, and, only one of REAL interest to us. There are exceptions (e.g., quantitative genetics provides many examples), but we are generally most interested in the among group test — this is the test of the main factor. In our example, the main factor was DRUG and whether the drugs differed in their effects on cholesterol levels. The second test is important in the sense that we prefer that it contributes little or no variation to the differences in cholesterol levels. But it might.

Table 14.5.6. F statistics for nested ANOVA.

F for the main effect is given as	$F=\frac{Groups\ MS}{Subgroups\ MS}$
F for the subgroup is given by	$F=\frac{Subgroups\ MS}{Error\ MS}$
and of course, use the appropriate DF when testing the F values!! The Critical Value \(F_{0.05(2), df \ numerator, df \ denominator\)	

One way to look at this: it would not make sense to conclude that an effect of the main group was significant if the variation in the subgroups was much, much larger. That's in part why we test the main effect with the subgroups MS and not the error MS. If variation due to the nested variable is not significant, then it is an estimate of the error variance, too.

The nested model we are describing is a two-factor ANOVA, but it is incomplete (compared to the balanced, fully crossed 2-way design we've talked about before). We don't have scores in every cell. Instead, each level of nested factor is paired with one and only one level of the other factor. In our example, Source is paired with only one other level of the other factor Drug (e.g., Source 1 goes with Drug 1 only), but the main effect is paired with 2 levels of the nesting factor (e.g., Drug 1 is manufactured at Source 1 and Source 2).

Note:

Nesting is strictly one-way. Drug is not nested within Source, for example.

Some important points about testing the null hypotheses in a nested design. For one, the test of the effect of the nesting factor (Source) is confounded by the interaction between the main factor. We don't actually know if the interaction is present, but we also get no way to test for it because of the incomplete design. We must therefore be cautious in our interpretation of the effect of the nested factor.

Consider our example. We want to interpret the effect of source as the contribution to the response based on variation among the different suppliers of the drugs. It might be good to know that some drug manufacturer is better (or worse) than others. However, differences among the sources for the different drugs are completely contained in the main effect factor (the test of effects of the different drugs themselves on the response). Therefore, the observed differences between sources COULD be entirely due to the effects of the different drugs and have nothing to do with variation among sources!!

Questions

1. Identify the response variable and whether the described factor (in all caps) is suitable for crossed design or nested design
   a. In a breeding colony of lab mice, BREEDERS are used to generate up to five LITTERS; effects on offspring REPRODUCTIVE SUCCESS.
   b. Effects of individual TEACHERS at different SCHOOLS on STUDENT LEARNING in biology.

c. Lisinopril, an ACE-inhibitor drug prescribed for treatment of high blood pressure, is now a generic drug, meaning a number of COMPANIES can manufacture and distribute the medication. Millions of DOSES of lisinopril are made each year; drug companies are required by the FDA to record when a dose is made and to record these dates by LOT NUMBER.

2. Work the example data set provide in this page. After loading the data set into Rcmdr (R), use linear model. The command to nest requires use of the forward slash, /. For example, if factor b is nested within factor a, then a/b. The linear model formula then,

```
Model <- lm(Obs ~ a/b, data=source)
```

1. Describe the problem, i.e., what is a? What is b? What are the hypotheses?
2. What is the statistical model?
3. Test the model.
4. Conclusions?

Data set used in this page

Drug	Source	Obs
A	s1	202.6
A	s1	207.8
A	s1	190.2
A	s1	211.7
A	s1	201.5
A	s2	189.3
A	s2	198.5
A	s2	208.4
A	s2	205.3
A	s2	210
B	s3	212.3
B	s3	204.4
B	s3	221.6
B	s3	209.2
B	s3	222.1
B	s4	203.6
B	s4	209.8
B	s4	204.1
B	s4	201.8
B	s4	202.6
C	s5	189.1
C	s5	219.9
C	s5	196
C	s5	205.3

C	s5	204
C	s6	194.7
C	s6	192.8
C	s6	226.5
C	s6	200.9
C	s6	219.7

14.6: Some other ANOVA designs

Introduction

There are several additional ANOVA models in common use. The crossed, balanced design is but one example of the two-way ANOVA. And, from a consideration of two factors, it logically follows that there can be more than two factors as part of the design of an experiment. As the number of factors increase, the number of two-way, three-way, and even higher-order interactions are possible and at least in principle may be estimated.

Our purpose here is to highlight several, but certainly not all, possible experimental designs from the perspective of ANOVA. Examples are provided. Keep in mind that the **general linear model** approach unifies these designs.

Some of the classical experimental ANOVA designs one sees include:

- Two-way randomized complete block design
- Two-way factorial with no replication design
- Repeat-measures ANOVA with one factor
- Nested ANOVA
- Three-way ANOVA
- Split-plot ANOVA
- Latin squares ANOVA

Put simply, these designs differ in how the groups are arranged and how members of the groups are included.

Two-way randomized complete block design

This design refers to the “textbook” design. For each, factor A and factor B, there are multiple levels, in this example three levels of Factor A and three levels of Factor B, and subjects (sampling units) are randomly assigned to each level. However, one of the factors is, perhaps, of less interest, yet certainly accounts for variation in the response variable.

		Factor A		
		1	2	3
Factor B	1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$
	2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$
	3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$

where $n_{1,1}$, $n_{2,1}$, etc. represents the number of subjects in each cell. Thus, in this design there are nine groups. Typically, minimum replication would be three subjects per group.

Two-way factorial with no replication design

While it may seem obvious that a good experiment should have replication, there are situations in which replication is impossible. While this seems rather odd, this scenario very much describes a typical microarray, gene expression project.

		Factor A		
		1	2	3
Factor B	1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$
	2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$
	3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$

where, again, $n_{1,1}$, $n_{2,1}$, etc., represents the number of subjects in each cell and there are nine groups in this study. With no replication, then there is no more than one subject per group.

Repeated-measures ANOVA

When subjects in the study are measured multiple times for the dependent variable, this is called a repeated-measures design. We introduced the design for the simple case of before and after measures on the same individuals in Chapter 12.3. It's straight-forward to extend the design concept to more than two measures on the subjects. The **blocking effect** is the individual (see Chapter 14.4), and, therefore, a random effect (see Chapters 12.3 and 14.3) in this type of experimental design.

Although straightforward in concept, repeated measure designs have many complications in practice. For example, long-term studies can expect for subjects to drop out of the study, resulting in **censored data**. Another complication affects the assumption is that there is no **carry over effect** — it doesn't matter the order different treatments are applied to the subjects. Think of this assumption as akin to the **equal variances assumption** in ANOVA; just like unequal variances effects Type I error rates in ANOVA, deviations from **sphericity** inflate Type I error rates in repeated-measures designs.

Sphericity assumption is described in two ways:

> **Assumption of sphericity** — the ranking of individuals remains the same across treatment levels — no interaction between individual and treatment. Sphericity assumption is always met if there are just two levels of the repeated measure, e.g., before and after.
>
> **Compound symmetry assumption** — the variances and covariances are equal across the study: the changes experienced by the subjects are the same across the study regardless of the order of treatments.

Tests for sphericity include:

Mauchly test: `mauchly.test(object)`

If results of tests for violations of sphericity warrant, corrections are available. One recommended correction is called **Greenhouse-Giesser correction**, which adjusts the degrees of freedom and so results in a better p-value estimate. A second correction is called **Huyhn-Feldt correction**; this correction, too, adjusts the degrees of freedom to improve the p-value estimate.

Three-way ANOVA

It is relatively straightforward to imagine an experiment that involves three or more factors. The analysis and interpretation of such designs, while feasible, becomes somewhat complicated, especially for the mixed models (Model III).

Consider just the case of a fixed-effects 3-way ANOVA. How many tests of null hypotheses are there?

1. Three tests for main effects.
2. Three tests of two-way interactions.
3. A test for a three-way interaction.

Thus, there are seven separate null hypotheses from a three-way ANOVA with fixed effects! As you can imagine, large sample sizes are needed for such designs, and the "higher-order" interactions (e.g., three-way interaction) can be difficult to interpret and may lack biological significance.

ANOVA designs without random assignment to treatment levels

Latin square design

We have introduced you to several ANOVA experimental designs that employed randomization for assignment of subjects to treatment groups. The purpose of randomization is even out differences due to confounding variables. However, if we know in advance something about the direction of the influence of these confounding variables, strictly random assignment is not in fact the best design. For example, the **Latin square** design is common in agriculture research and is very useful for situations in which two gradients are present (e.g., soil moisture levels, soil nutrient levels).

	Dry soil ← → Wet soil			
Soil Nutrients	T1	T4	T3	T2
	T3	T2	T1	T4
low ↑ ↓	T2	T3	T4	T1

high	T4	T1	T2	T3

Split-Plot Design

Another design from agriculture research is especially useful to ecotoxicology research. We mentioned the repeated measures design in which individuals are measured more than once and each individual receives all levels of the treatment in a random order (**cross-over design**). However, this design assumes that there are no **carry-over effects** (see Hills and Armitage 1979; For ecology/evolution definition see O'Connor et al 2014). While this assumption may hold for many experiments, we can also imagine many more situations in which this is undoubtedly false. For example, if we wish to measure the effects of ozone and relative humidity on frog behavior, we might consider using the individual as its own control. But we also wish to compare frog behavior following ozone exposure against behavior exhibited in clean air. But we are likely to violate the carry-over assumption. If a frog receives ozone then air, the effects of ozone may inhibit activity for several days after the initial exposure, which would then influence subsequent measures. The solution to this dilemma is to use what's called a **split plot** design. The design combines elements of nesting.

Consider our frog experiment. There would be three factors:

Factor 1 = Exposure (air or ozone),

Factor 2 = Saturation (dry, intermediate, wet),

Factor 3 = Individual (each frog is measured 3 times).

The design table would look like

		Exposure					
		Air			**Ozone**		
Humidity	**Dry**	Frog1	Frog2	Frog3	Frog4	Frog5	Frog6
	Intermediate	Frog1	Frog2	Frog3	Frog4	Frog5	Frog6
	Wet	Frog1	Frog2	Frog3	Frog4	Frog5	Frog6

Thus, the design is crossed for one factor (saturation), but nested for another factor (individuals are nested within Exposure factor).

Questions

1. Which of the study designs mentioned so far are sensitive to carry-over effects?
2. With respect to how levels of Factors are assigned, distinguish the split-plot design from the Latin square design.

14.7: Rcmdr Multiway ANOVA

Introduction

We have been talking about the two-way randomized, balanced, replicated design. Here, we take you step by step through use of R to conduct the **multiway ANOVA**.

R code: Multiway ANOVA

Rcmdr: Statistics → Means → Multiway ANOVA… we will review this as Option 1

or

Rcmdr: Statistics → Fit Models → Linear model… we will review this as Option 2

In either case, as a reminder, your data set must be a stacked worksheet, like the data in this table.

Table 14.7.1. Data set, example.14.7†

Diet	Population	Response
A	1	4
A	1	6
A	1	5
A	2	5
A	2	8
A	2	9
B	1	12
B	1	15
B	1	11
B	2	5
B	2	7
B	2	8

Option 1

Your first option is to use the ANOVA menus via "Means." This is a perfectly good way to handle a standard two-way, fully-crossed, **fixed effects** model. However, other designs will not run with this command and R will return a report of errors for ANOVA models that do not conform to the replicated, balanced, crossed design.

Rcmdr: Statistics → Means → Multiway Analysis of variance …

Factors: highlight "Diet" AND "Population"

Response variable: pick one (in this window, all we see is "Response")

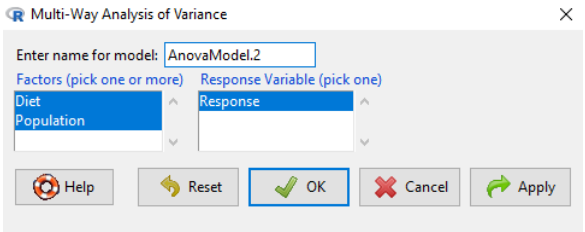


Figure 14.7.1: Multi-Way Analysis of Variance popup in R Commander.

†Note:

Don’t forget to convert numeric Population to factor

Interpret the output

```
AnovaModel.2 <- (lm(Response ~ Diet*Population, data=example.14.7))

Anova(AnovaModel.2)

Anova Table (Type II tests)
Response: Response
                  Sum Sq    Df    F value      Pr(>F)
Diet              36.750     1    12.2500    0.008079 **
Population        10.083     1     3.3611    0.104104
Diet:Population   52.083     1    17.3611    0.003136 **
Residuals         24.000     8
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
tapply(example.14.7 Response, list(Diet=example.14.7 Diet,
+ Population=example.14.7 Population), mean, na.rm=TRUE) # means
Population
Diet 1 2
A 5.00000 7.333333
B 12.66667 6.666667

tapply(example.14.7 Response, list(Diet=example.14.7 Diet,
+ Population=example.14.7 Population), sd, na.rm=TRUE) # std. deviations
Population
Diet 1 2
A 1.000000 2.081666
B 2.081666 1.527525

tapply(example.14.7 Response, list(Diet=example.14.71 Diet,
+ Population=example.14.7 Population), function(x) sum(!is.na(x))) # counts
Population
Diet 1 2
A 3 3
B 3 3
```

End R output

Summary of multi-way ANOVA command

The multi-way ANOVA command returns our ANOVA table plus the **adjusted means**, along with **standard deviations** and number of observations (counts). The adjusted means would then be good to put into a chart to present group comparisons following adjustments from the effects of levels within groups.

Rcmdr: Models → Graphs → Predictor effect plots …

Here's the chart (hint: $\pm SEM = \frac{SD}{\sqrt{count}}$)

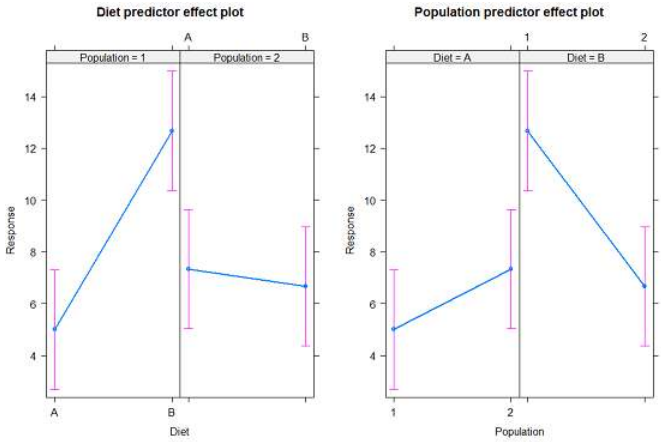


Figure 14.7.2: Plots of the predictor effects of each factor.

Option 2

A more general approach is to use the General linear model. This approach can handle the standard 2-way fixed effects ANOVA (above), but any other model as well. The model is `Response ~ Diet*Population` .

Rcmdr: Statistics → Fit Models → Linear model…

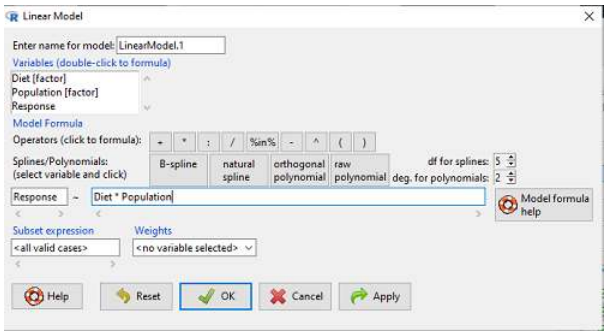


Figure 14.7.3: Linear Model screenshot in R Commander with model formula input.

Interpret the output

```
LinearModel.1 <- lm(Response ~ Diet * Population, data=example.14.7)

summary(LinearModel.1)

Call:
lm(formula = Response ~ Diet * Population, data = example.14.7)

Residuals:
Min 1Q Median 3Q Max
-2.3333 -1.1667 0.1667 1.0833 2.3333

Coefficients:
                        Estimate    Std. Error ..t value .  Pr(>|t|)
(Intercept) .              5.000         1.000 .    5.000    .0.00105 **
Diet[T.B] .                7.667       ..1.414 .    5.421    .0.00063 ***
Population[T.2]           .2.333       ..1.414 .    1.650    .0.13757
```

```
Diet[T.B]:Population[T.2] -8.333       ..2.000    -4.167    .0.00314 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.732 on 8 degrees of freedom
Multiple R-squared: 0.8047, Adjusted R-squared: 0.7315
F-statistic: 10.99 on 3 and 8 DF, p-value: 0.003285
```

End R output

Lots to sort through, so let's begin with what is in common between the two approaches in `Rcmdr`, the Multi-way ANOVA command versus the linear model command.

Compare the two outputs

As a direct output, the linear model option does not provide an ANOVA summary table. Instead of our ANOVA table, the linear model returns estimates of coefficients along with t-test results for each coefficient of the model from the `lm()` command output

Recall that we can get ANOVA tables through the following R commands via `Rcmdr`.

Rcmdr: Models → Hypothesis tests → ANOVA Table.

Let's do so for this linear model (accept the default for type of tests = "Type II").

And the output is

```
Anova(LinearModel.1, type="II")
Anova Table (Type II tests)

Response: Response
               Sum Sq Df F value   Pr(>F)
Diet           36.750  1 12.2500 0.008079 **
Population     10.083  1  3.3611 0.104104
Diet:Population 52.083  1 17.3611 0.003136 **
Residuals      24.000  8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

End R output

Now we're in business, and, using the `lm()` function, we have the estimates for each model coefficient plus our ANOVA table.

Both methods give the same answer! Of course. Which to choose, Option 1 or Option 2? Use the `lm()` option: it is more flexible and covers more designs than the multiway ANOVA, which is strictly for the crossed fully replicated design.

Questions

1. Write out the two-way model described for the data in Table 14.7.1.
2. Write the null hypotheses and provide a summary of the statistical significance of the model.

14.8: More on the linear model in rcmdr

Introduction

During the last lectures, we could have used the Two-Way ANOVA command in R or Rcmdr.

R code: How to analyze multifactorial ANOVA problems

Rcmdr: Statistics → Means → Multiway ANOVA

To analyze our two-factor data sets. As long as the design meets the following conditions, by all means use this command because it is simple and precisely correct.

- Both factors are fixed, not random.
- Each level of first factor is crossed with each level of the second factor.
- No missing data (the design is fully replicated and balanced).

If any of the three points do not fit your two-way design, then you'll need a different, more general and powerful ANOVA procedure in R and Rcmdr to analyze these types of designs. You'll need the `lm()` function (Fig. 14.8.1).

Rcmdr: Statistics → Fit models → Linear model…

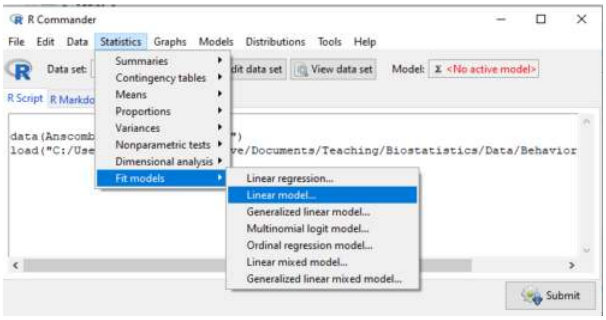


Figure 14.8.1: *Linear model menu in Rcmdr, version 2.7.0*

Some R basics with the `lm()` function, the general linear model. $Y \sim model$

Response Y is specified by linear predictor(s), either factors or covariates (ratio-scale predictor variables). We communicate to R what the model is by using operators. The four most commonly used operators are:

$+$	the basic way to include the model terms, i.e., the mode predictors
$-$	which is interpreted as the interaction of all the variables and the factors in the term
$*$	which is interpreted as factor crossing
$\%in\%$	indicates the term on the left is nested within the term on the right.

A few examples: we'll have three factors, ***A***, ***B***, and ***C***. For our one-way ANOVA, the model specification would be $Y \sim A$

For our crossed, balanced two-way ANOVA, the model specification would be $Y \sim A + B + A : B$

or equivalently $Y \sim A * B$

And for our block ANOVA problem?

Click here to get the entire list of model commands in R.

How would we analyze our snake experiment?

Table 14.8.1. The snake data set

Snake	Source	flick
1	dH2O	3
2	dH2O	0
3	dH2O	0

4	dH2O	5
5	dH2O	1
6	dH2O	2
1	fish	6
2	fish	22
3	fish	12
4	fish	24
5	fish	16
6	fish	16
1	worm	6
2	worm	22
3	worm	12
4	worm	24
5	worm	16
6	worm	16

We have two factors, but one factor is a **Block** (repeated measures on individuals). We need to tell R and Rcmdr what our model is. We'll return to talk about models next time, a very important topic!! For now, think of a model as adding the independent variables together to predict the response variable.

In our Snake example, it's a two-way ANOVA, but one factor is individual snake, the other is a treatment, and we have repeat measures, so there cannot be an interaction.

We tell R and Rcmdr which columns contain the Response, and under Model, we enter the columns with the two factors.

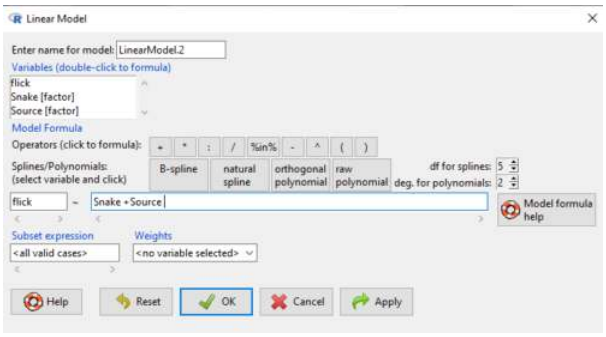


Figure 14.8.2: Menu of linear model with repeat measures model, Rcmdr, version 2.7.0.

You must also tell R and Rcmdr which factors in the model (if any) are random to get the correct F statistics. Almost without exception, blocking factors are always treated as Random

The output looks like this (see below). More complicated, true (which means more information!), but things marked in red we've seen before.

```
LinearModel.2 <- lm(flick ~ Snake +Source, data=L16SnakeTaste)

summary(LinearModel.2)

Call:
```

```
lm(formula = flick ~ Snake + Source, data = L16SnakeTaste)

Residuals:
Min 1Q Median 3Q Max
-5.2222 -0.7222 0.0278 1.5694 7.4444

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.444 2.523 -1.762 0.10862
Snake[T.2] 9.667 3.090 3.128 0.01072 *
Snake[T.3] 3.000 3.090 0.971 0.35451
Snake[T.4] 12.667 3.090 4.099 0.00215 **
Snake[T.5] 6.000 3.090 1.942 0.08085 .
Snake[T.6] 6.333 3.090 2.050 0.06755 .
Source[T.fish] 14.167 2.185 6.484 0.0000704 ***
Source[T.worm] 14.167 2.185 6.484 0.0000704 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.784 on 10 degrees of freedom
Multiple R-squared: 0.8858, Adjusted R-squared: 0.8058
F-statistic: 11.08 on 7 and 10 DF, p-value: 0.0005337
```

For the ANOVA table, we call up the command via

Rcmdr: Models → Hypothesis tests → ANOVA table… (Fig. 14.8.3).

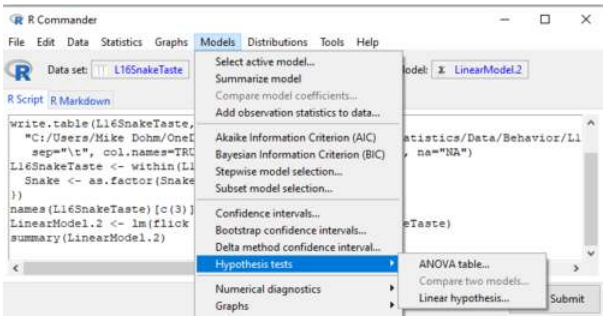


Figure 14.8.3: Rcmdr: Models → Hypothesis tests → ANOVA table… Rcmdr, version 2.7.0

Confirm that the model object is active (in this case, the object was LinearModel.2), accept the defaults about types of tests and marginality, and submit OK. The output is

```
Anova(LinearModel.2, type="II")
Anova Table (Type II tests)

Response: flick
Sum Sq Df F value Pr(>F)
Snake 307.61 5 4.2956 0.02396 *
Source 802.78 2 28.0256 0.00007954 ***
Residuals 143.22 10
```

```
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

End R output

You do not need to know all of the output; all of that out put *is* there for a reason, of course, but for now, here's what R and Rcmdr has to say (from the help menu):

The sequential sums of squares is the **added sums of squares** given that prior terms are in the model. These values depend upon the model order. The **adjusted sums of squares** are the sums of squares given that all other terms are in the model. These values do not depend upon the model order.

You should try our snake example again, but this time, remove the tongue flick response to the dH2O for the first snake — a missing value. (just type into the cell NA).

How would you use GLM to analyze a simple two-way ANOVA, a fully crossed, fully replicated ("balanced") design? We could use the Two-way ANOVA command in R and Rcmdr, or we could use `lm()` . Try it with the data set from the lecture on random vs nonrandom.

Another example

Below, you see how the model is entered. Note to indicate that I wish R and Rcmdr to test the **interaction**, I need to add a Model term for that source of variation. I accomplish this by typing "Diet*Drug" (without the quotes).

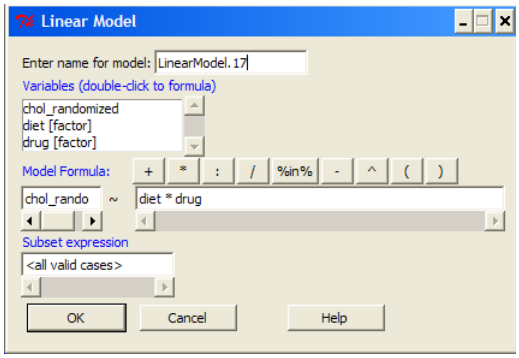


Figure 14.8.4: Crossed, balanced design. Linear model menu, Rcmdr, version 1.9.2.

After clicking OK, the following output from the `lm` function is returned. How does this compare to output from the two-way ANOVA command in R and Rcmdr?

You should try both and compare!

Rcmdr: Models → Hypothesis tests → ANOVA table

```
Anova(LinearModel.17, type="II")
Anova Table (Type II tests)
Response: chol_randomized
.............Sum Sq ..Df .F value ...Pr(>F)
diet ........141.08 ...2 ..1.3061 ..0.28745
drug ........351.88 ...2 ..3.2577 ..0.05403 .
diet:drug ...235.50 ...4 ..1.0901 ..0.38124
Residuals ..1458.19 ..27
```

End R output

Nested ANOVA

The nested ANOVA may be analyzed in multiple ways in R and Rcmdr, but I prefer the `lm()` function because it is the most general. For Nested ANOVA, we can also use `lm()` . Here's where it gets a little tricky. Put in the Response variable (Chol), then

click in the box for model: Select both factors, then type in / after the factor that's nesting factor. For our nested model example (14.5 – Nested designs), Manufacturer Source was nested within Drug.

Table 14.8.3. Nested design example data set from Chapter 14.5 – Nested design

Drug	Source	Chol
1	1	202.6
1	1	207.8
1	1	190.2
1	1	211.7
1	1	201.5
1	2	189.3
1	2	198.5
1	2	208.4
1	2	205.3
1	2	210
2	3	212.3
2	3	204.4
2	3	221.6
2	3	209.2
2	3	222.1
2	4	203.6
2	4	209.8
2	4	204.1
2	4	201.8
2	4	202.6
3	5	189.1
3	5	219.9
3	5	196
3	5	205.3
3	5	204
3	6	194.7
3	6	192.8
3	6	226.5
3	6	200.9
3	6	219.7

Note that if you were working with a CROSSED model, then you would enter the two factors and indicate the interaction by typing Drug*Source (if these are the two factors involved in the interaction).

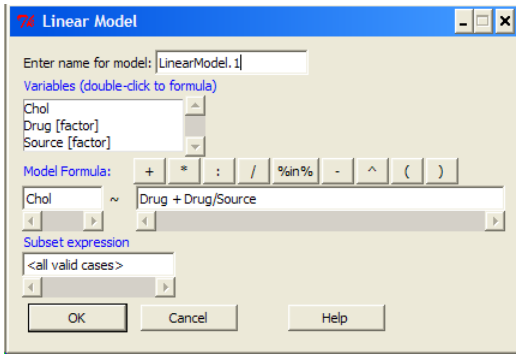


Figure 14.8.5: Nested design, linear model menu, Rcmdr, version 1.9.2.

Fortunately, R and Rcmdr's help system is quite extensive here, so when in doubt, check the help box…

Output from the linear model for the Nested Example looks like the one below.

The General Linear Model function in R and therefore Rcmdr returns information about our design plus Sums of Squares, Mean squares, and P-values. R and Rcmdr default's to use of sequential evaluation of effects. Adjusted evaluation is useful for when you have a covariate (like body size or another confounding variable) that should be evaluated first before the factors are evaluated. We will use the sequential analysis.

Rcmdr: Models → Hypothesis tests → ANOVA table

```
Anova(LinearModel.15, type="II")
Anova Table (Type II tests)
Response: Chol
...............Sum Sq ..Df ...F value ..Pr(>F)
Drug ..........225.14 ...2 ....1.1743 ..0.3262
Drug:Source ...269.27 ...3 ....0.9363 ..0.4385
Residuals ....2300.61 ..24
```

End R output

Repeatability and ANOVA

We need to tell Rcmdr how to structure the error term; you need the data frame to be arranged so

```
aovRes <- aov(dH2O ~ Source + Error(Source/Subject), data=SnakeTaste)
#Print the results
aovRes
Anova Table (Type II tests)
Response: dH2O
           Sum Sq Df F value     Pr(>F)
Subject 307.61  5  4.2956    0.02396 *
Source      802.78  2 28.0256 0.00007954 ***
Residuals   143.22 10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#What components are available in the aovRes object?
names(aovRes)
[1] "Sum Sq"  "Df"      "F value" "Pr(>F)"
```

How do I extract the "F value" for Subjects?

```
str(aovRes)
```

Questions

[pending]

14.9: Chapter 14 References

Hills, M., & Armitage, P. (1979). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology*, *8*(1), 7–20.

Hills, M., & Armitage, P. (2004). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology*, *58*(7), S717–S719.

Meijer, J. H., & Robbers, Y. (2014). Wheel running in the wild. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1786), 20140210.

Mills, E. J., Chan, A.-W., Wu, P., Vail, A., Guyatt, G. H., & Altman, D. G. (2009). Design, analysis, and presentation of crossover trials. Trials, 10(1), 27. https://doi.org/10.1186/1745-6215-10-27

National Research Council. (2005). *Mathematics and 21st century biology*. National Academies Press.

O'Connor, C. M., Norris, D. R., Crossin, G. T., & Cooke, S. J. (2014). Biological carryover effects: Linking common concepts and mechanisms in ecology and evolution. Ecosphere, 5(3), art28. https://doi.org/10.1890/ES13-00388.1

See references in Chapter 12.

CHAPTER OVERVIEW

15: Nonparametric Tests

Introduction

t-tests and ANOVA are members of a statistical family of tests called **parametric tests.** Parametric tests assume that the

- sample of observations come from a particular probability distribution, e.g., normal distribution.
- samples among groups have equal variances.

Assumptions for parametric tests were introduced in Chapter 13. In the case of the t-test and ANOVA, we assume that the samples come from a **normal probability distribution** and that the probabilities of the test statistic follow the t distribution or the F distribution, respectively.

Providing these assumptions hold, we can then proceed to interpret our results as if we are talking about the population as a whole from which the samples were selected.

In other words, the t-test asks (infers) about properties of a population; hence, we are asking about parameters of the population.

t-tests, ANOVA, and other parametric tests are designed to work with quantitative ratio-scale data types. If the data are of this type and the probability distribution is known, they are the best tests to use… they allow you to make conclusions about experiments at a defined **Type I error rate** = 5%.

But what if you can't assume that distribution? Your options include

- transforming the data so as the data better meet the assumptions of parametric tests.
- apply nonparametric statistical tests.

That's where **nonparametric** statistics come in as an alternative to parametric tests.

Nonparametric tests make fewer assumptions

Nonparametric tests do not make the assumption about a particular distribution — **distribution-free tests** — nor are they used to make inferences about population parameters. Instead, nonparametric tests are used when the data type are ranks (ordinal). Now, when you think about it, all quantitative data can be converted to ranks. Hence, this is the argument for why there are nonparametric alternatives for tests like the t-test. There are a number of nonparametric alternatives to parametric tests. Another nonparametric option is to run a **permutation test** on the data.

Nonparametric tests lack statistical power

One downside for nonparametric tests is that they tend to have less statistical power compared to the parametric alternatives (see Chapter 11 for a review of **Statistical Power**). Thus, nonparametric tests tend to have higher Type II rates of error — they fail to properly reject the null hypothesis when they should. This problem tends to be less important for large sample sizes.

> **Note:**
>
> Instead of **transformations** or other ad hoc manipulations of the data, modern statistical approaches favor modeling the error structure of the data within a Generalized Linear Model framework (St.-Pierre et al 2018). The advantage of the model approach is that parameter estimation occurs on the raw data. Use of transformations may, however, remain a better choice. While statistically justified, the generalized linear model approach may also tend to have higher rates of Type II error compared to simple transformations.

Thus, this chapter covers some of the more popular nonparametric alternative tests. Chapter 19 – Distribution-free statistical methods highlights use of permutation and randomization approaches, which also are alternatives to parametric tests.

15.1: Kruskal-Wallis and ANOVA by ranks

Introduction

When the data are NOT normally distributed OR when the variances in the different samples are NOT equal, one option is to opt for a non-parametric alternative and use the **Kruskal-Wallis test**.

It is known that the an **ANOVA on ranks** of the original data will yield the same results as the original data.

Kruskal-Wallis

Rcmdr: Statistics → Nonparametric test → Kruskal-Wallis test…

Rcmdr Output of Kruskal-Wallis test

```
tapply(Pop_data$Stuff, Pop_data$Pop, median, na.rm=TRUE)

Pop1 Pop2 Pop3 Pop4
 146   90  122  347
kruskal.test(Stuff ~ Pop, data=Pop_data)
Kruskal-Wallis rank sum test
data: Stuff by Pop
Kruskal-Wallis chi-squared = 25.6048, df = 3, p-value = 1.154e-05
```

End of R output

So, we reject the **null hypothesis**, right?

Compare parametric test and alternative non-parametric test

Let's compare the nonparametric test results to those from an analysis of the ranks (ANOVA of ranks).

To get the ranks in R Commander (example.15.1 data set is available at bottom of this page; scroll down or click here).

Rcmdr: Data → Manage variables in active data set → Compute new variable …

The command for ranks is…. wait for it …. `rank()` . In the popup menu box, name the new variable (`Ranks`) and in the Expression to compute box enter `rank(Values).`

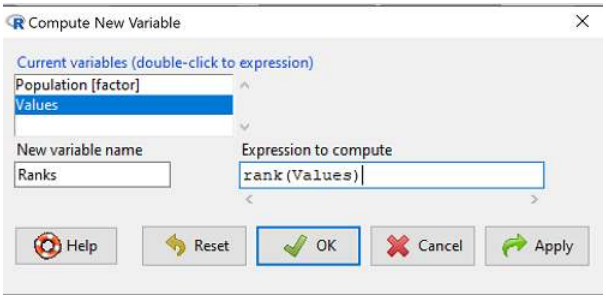


Figure 15.1.1: Screenshot of Rcmdr menu, Create New Variable.

Note:

It's not a good idea to name an object `Ranks` , because that's similar to a function name in R, `rank` .

And the R code is simply

```
example.15.1$Ranks <- with(example.15.1, rank(Values))
```

Note:

The object `example.15.1$Ranks` adds our new variable to our data frame.

That's one option, to rank across the entire data set. Another option would be to rank within groups.

R code:

```
example.15.1$xRanks <- ave(Values, Population, FUN=rank)
```

Note:

The `ave()` function averages within subsets of the data and applies whatever summary function (`FUN`) you choose. In this case we used `rank` . Alternative approaches could use split or `lapply` or variations of `dplyr` . `ave()` is in the base package and at least in this case is simply to use to solve our rank within groups problem. `xRanks` would then be added to the existing data frame.

Here are the results of ranking within groups.

Population 1	Rank1	Population 2	Rank2	Population 3	Rank3	Population 4	Rank4
105	11.5	100	9	130	17.5	310	33
132	19	65	4	95	7	302	32
156	22	60	2.5	100	9	406	38
198	29	125	16	124	15	325	34
120	13.5	80	5.5	120	13.5	298	31
196	28	140	21	180	26	412	39.5
175	24	50	1	80	5.5	385	39.5
180	26	180	26	210	30	329	35
136	20	60	2.5	100	9	375	37
105	115	130	17.5	170	23	365	36

Question. Which do you choose, rank across groups (`Ranks`) or rank within groups (`xRanks`)? Recall that this example began with a nonparametric alternative to the one-way ANOVA, and we were testing the null hypothesis that the group means were the same.

$H_O : \bar{X}_1 = \bar{X}_2 = \bar{X}_3$

Answer. Rank the entire data set, ignoring the groups. The null hypothesis here is that there is no difference in **median ranks** among the groups. Ranking within groups simply shuffles observations within the group. This is basically the same thing as running Kruskal-Wallis test.

Run the one-way ANOVA, now on the Ranked variable. The ANOVA table is summarized below.

Source	DF	SS	MS	F	P†
Population	3	3495.1	1165.0	22.94	< 0.001
Error	36	1828.0	50.8		
Total	39	5323.0			

Note:

† The **exact p-value** returned by R was 0.0000000178. This level of precision is a bit suspect given that calculations of p-values are subject to bias too, like any estimate. Thus, some advocate to report p-values to three significant figures, and if less than 0.001, report as shown in this table. Occasionally, you may see P = 0.000 written in a journal article. This is a definite no-no; p-values are estimates of the probability of getting results more extreme then our results and the null hypothesis holds. It's an estimate, not certainty; p-values cannot equal zero.

So, how do you choose between the parametric ANOVA and the non-parametric Kruskal-Wallis (ANOVA by Ranks) test? Think like a statistician — It is all about the type I error rate and potential bias of a statistical test. The purpose of statistics is to help us separate real effects from random chance differences. If we are employing the **NHST** approach, then we must consider our chance that we are committing either a Type I error or a Type II error, and conservative tests, e.g., tests based on comparing medians and not means, implies an increased chance of committing **Type II errors**.

Questions

1. Saying that nonparametric tests make fewer assumptions about the data should not be interpreted that they make no assumptions about the data. Thinking about our discussions about experimental design and our discussion about test assumptions, what assumptions must hold regardless of the statistical test used?
2. Go ahead and carry out the one-way ANOVA on the within group ranks (`xRanks`). What's the p-value from the ANOVA?
3. One could take the position that only nonparametric alternative tests should be employed in place of parametric tests, in part because they make fewer assumptions about the data. Why is this position unwarranted?

Data used in this page

Dataset for Kruskal-Wallis test

Population	Values
Pop1	105
Pop1	132
Pop1	156
Pop1	198
Pop1	120
Pop1	196
Pop1	175
Pop1	180
Pop1	136
Pop1	105
Pop2	100
Pop2	65
Pop2	60
Pop2	125
Pop2	80
Pop2	140
Pop2	50
Pop2	180

Population	Values
Pop2	60
Pop2	130
Pop3	130
Pop3	95
Pop3	100
Pop3	124
Pop3	120
Pop3	180
Pop3	80
Pop3	210
Pop3	100
Pop3	170
Pop4	310
Pop4	302
Pop4	406
Pop4	325
Pop4	298
Pop4	412
Pop4	385
Pop4	329
Pop4	375
Pop4	365

Simulated values from three populations

15.2: Wilcoxon rank sum test

Introduction

Wilcoxon rank sum test, also called the **two-sample Wilcoxon test**, is a nonparametric test. It is equivalent to another nonparametric test called the **Mann-Whitney test**, which was independently derived. We get the Wilcoxon test statistic in `Rcmdr` through the Statistics submenu.

Rcmdr: Statistics → Nonparametric tests → Two-sample Wilcoxon Test

I'll show you the test with an example. We'll use the same data set introduced in chapter 10.3, body mass (g) for four **geckos** (*Hemidactylus frenatus*, Fig. 15.2.1) and four green **anolis lizards** (*Anolis carolinensis*, Fig. 15.2.2).

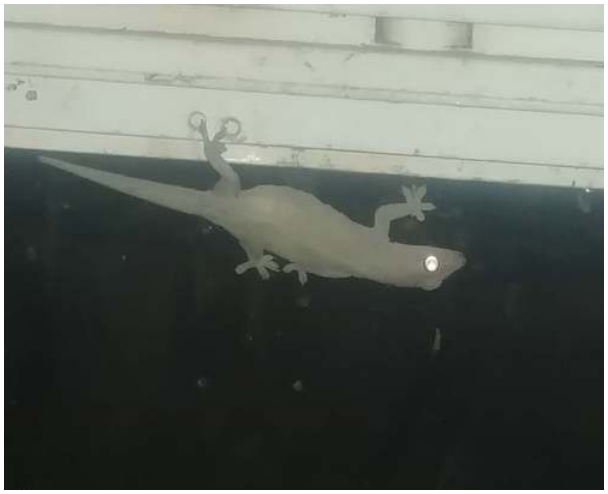


Figure 15.2.1: Female common house gecko, *Hemidactylus frenatus*, central Oahu, M. Dohm 2018.

Figure 15.2.2: Male *Anolis carolinensis*, 'Akaka Falls, Hawai`i, M. Dohm 2018.

Wilcoxon test, worked example

```
Geckos: 3.186, 2.427, 4.031, 1.995
Anoles: 5.515, 5.659, 6.739, 3.184
```

Note:

This test in `Rcmdr` requires that data are in a **stacked worksheet** and not in **unstacked worksheet** with two columns. If you need help with worksheet format, then see Part07 in Mike's Workbook for Biostatistics.

We choose from the Rcmdr Nonparametric statistics menu the Two-sample Wilcoxon test (Fig. 15.2.3), then a **two-tailed test** of the **null hypothesis** (Fig. 15.2.4) and elect to use the defaults for the tests and calculations of P-values.

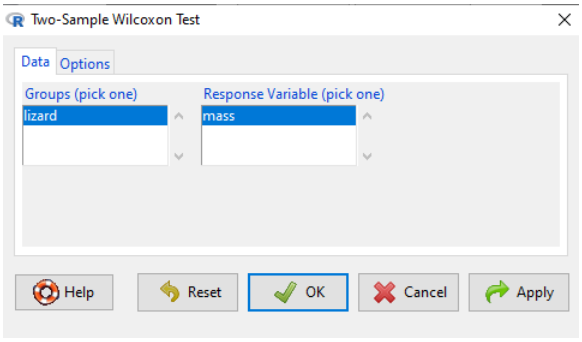


Figure 15.2.3: Screenshot Rcmdr menu 2 sample Wilcoxon test. Options are selected by clicking on "Options" tab (see Fig. 15.2.4)

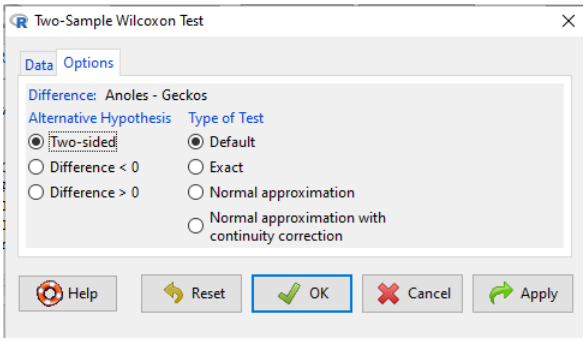


Figure 15.2.4: Screenshot of options tab Rcmdr menu 2 sample Wilcoxon test. Keep defaults to run the "Wilcoxon test."

Don't forget to stack the data. Rcmdr won't produce an error message if the data set is in the unstacked, improper conformation. Instead, Rcmdr menu options will not be available. For example, Fig. 15.2.5 shows a Two-sample Wilcoxon test… dimmed from view, not available for selection.

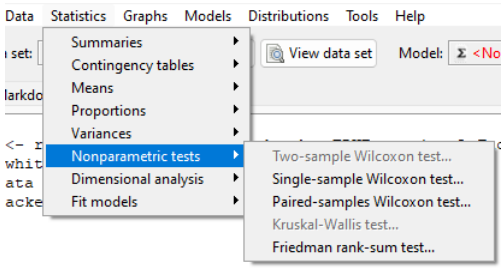


Figure 15.2.5: Screenshot of Rcmdr menu. Note Two-sample Wilcoxon test… is not available.

The results of the test, copied from the Output window, are shown below.

```
wilcox.test(Mass ~ Lizard, alternative="two.sided", data=LizardStacked)

Wilcoxon rank sum test
data: Mass by Lizard
W = 14, p-value = 0.1143
alternative hypothesis: true location shift is not equal to 0
```

The calculation of the **Wilcoxon test statistic (W)** is straightforward, involving summing the ranks. Obtaining the p-value of the test of the null is a bit more involved as it depends on permutations of all possible combinations of differences. For us, R will do nicely with the details, and we just need to check the p-value.

Here, we see that the medians are 5.6 g for the *Anolis*, and 2.8 g for the geckos. The associated p-value is 0.1143. Thus, we fail to reject the null hypothesis and conclude that there was no difference in median body mass. Note that this is the same general conclusion we got when we ran a independent t-test on this data set: there is no difference between day one and day two.

Questions

1. Conduct an independent t-test on the Lizard body mass data.
   - Make a box plot to display the two groups and describe the middle and variability.
   - Compare results of test of hypothesis. do they agree with the Wilcoxon test? If not, list possible reasons why the two tests disagree.
2. Using the dataset below, test null hypothesis using independent t-test, Welch's test, and nonparametric Wilcoxon's test.
   - Make a box plot to display the two groups and describe the middle and variability.
   - Compare results of test of hypothesis. do they agree with the Wilcoxon test? If not, list possible reasons why the tests disagree.

Data set

var1	var2
5.84	5.93
5.72	5.95
5.75	6.02
5.78	5.81
5.81	6.16
5.81	5.95
5.73	6.09
5.77	5.89
5.76	5.99
5.86	5.60
5.84	6.16
5.83	6.16
5.80	6.06
5.78	6.07
5.89	5.66
5.83	6.14
5.79	5.99
5.84	6.15
5.90	5.81
5.86	6.20

15.3: Wilcoxon signed-rank test

Introduction

When experimental units are repeated or paired, they lack independence and evaluating any difference between paired groups without accounting for the association between the repeat measures or between the pairs of related subjects would lead to incorrect inferences. A familiar pairing of experimental units occurs in clinical observational research in which control subjects and treatment subjects are matched by many characteristics.

In such cases, the parametric **paired t-test** would be used to evaluate inferences about the differences between repeat measures or between treatment and matched control subjects for some measured outcome. A nonparametric alternative to the paired t-test is the **Wilcoxon signed rank test**, also called the **paired Wilcoxon test**.

Another common example of paired sampling units would be that individuals are measured more than once for the same character or feature. For example, in Chapter 10.3 we presented results of running pace in minutes to complete the race of 15 women for repeated trials (in different years) at a 5K race held annually in Honolulu (Table 1).

Table 15.3.1. 5K repeat measures running data from Chapter 10.3.

ID	Race 1	Race 2
1	15.28	15.61
2	11.22	11.19
3	8.80	9.14
4	8.88	5.46
5	9.81	10.50
6	6.12	5.69
7	8.31	8.71
8	6.26	7.42
9	17.16	16.41
10	16.23	15.82
11	5.90	7.12
12	8.31	10.48
13	5.93	8.64
14	10.54	5.99
15	9.53	8.69

To get the paired Wilcoxon test in R Commander, select **Rcmdr: Statistics → Nonparametric tests → Paired Wilcoxon Test**

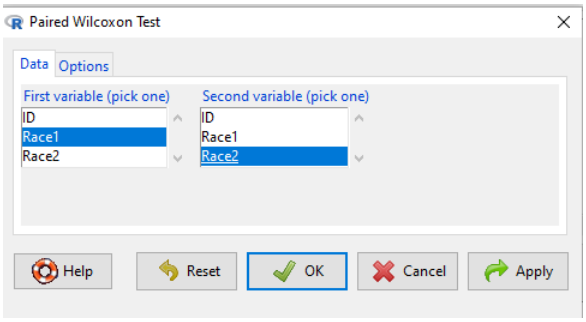


Figure 15.3.1: R Commander paired Wilcoxon test menu (aka Wilcoxon signed rank sum test). Rcmdr version 2.7.

Select first variable (e.g., Race1), second variable (e.g., Race2). Next, select Options tab and set null hypothesis. Accept defaults (Fig. 15.3.2).

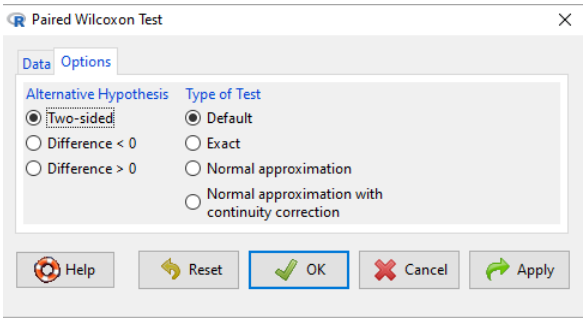


Figure 15.3.2: R Commander Options, select null hypothesis.

For the nonparametric paired Wilcoxon test we choose among the options to set our conditions; from the context menu a two-tailed test and we elect to use the defaults for the tests and calculations of p-values.

The results, copied from the Output window, are shown below. The calculation of the Wilcoxon test statistic (V) is straightforward, involving summing the ranks. Obtaining the p-value of the test of the null is a bit more involved as it depends on permutations of all possible combinations of differences. For us, R will do nicely with the details, and we just need to check the p-value.

```
with(repeat15_banana5K, wilcox.test(Race.1, Race.2, alternative='two.sided', paired=TI
# median difference
[1] -0.3313126

Wilcoxon signed rank exact test
data: Race.1 and Race.2
V = 58, p-value = 0.9341
alternative hypothesis: true location shift is not equal to 0
```

End R output

Here, we see that the median difference is small (-0.33), and the associated p-value is 0.93. Thus, we failed to reject the null hypothesis and should conclude that there was no difference in median running pace during the first and second trials.

Note that this is the same general conclusion we got when we ran a paired t-test on this data set: no difference between day one and day two.

R code

```
example.ch10.3 <- read.table(header=TRUE, text = "
ID Race1 Race2
1 15.28 15.61
2 11.22 11.19
3 8.80 9.14
4 8.88 5.46
5 9.81 10.50
6 6.12 5.69
7 8.31 8.71
8 6.26 7.42
9 17.16 16.41
10 16.23 15.82
11 5.90 7.12
```

```
12 8.31 10.48
13 5.93 8.64
14 10.54 5.99
15 9.53 8.69
")
```

Questions

1. This question lists all fourteen statistical tests we have been introduced to so far
   a. Mark **yes** or **no** as to whether or not the test is a parametric test
   b. Identify the nonparametric test(s) with their equivalent parametric test(s). If there are no equivalency, simply write "none."

	Parametric test? Yes/No	**If nonparametric, write the number(s) of the tests that the nonparametric test serves as an alternate for**
1. ANOVA by ranks		
2. Bartlett Test		
3. Chi-squared contingency table		
4. Chi-squared goodness of fit		
5. Fisher Exact test		
6. Independent-sample t-test		
7. Kruskal-Walis test		
8. Levene test		
9. One-sample t-test		
10. One-way ANOVA		
11. Paired t-test		
12. Shapiro-Wilks test		
13. Tukey post-hoc comparisons		
14. **Welch's test**		

15.4: Chapter 15 References and Suggested Reading

Bewick, V., Cheek, L., Ball, J. (2004). Statistics review 10: further nonparametric methods. *Critical Care* 8:196-199.

Fagerland, M. W., Sandvik, L. (2009). The Wilcoxon–Mann–Whitney test under scrutiny. *Statistics in Medicine* 28:1487-1497.

Feltovich, N. (2003). Nonparametric Tests of Differences in Medians: Comparison of the Wilcoxon–Mann–Whitney and Robust Rank-Order Tests. *Experimental Economics* 6:273-297

St-Pierre, A. P., Shikon, V., & Schneider, D. C. (2018). Count data in biology—Data transformation or model reformation?. *Ecology and evolution*, *8*(6), 3077-3085.

Whitley, E., Ball, J. (2002). Statistics review 6: Nonparametric methods. *Critical Care* 6:509-513.

CHAPTER OVERVIEW

16: Correlation, Similarity, and Distance

Introduction

We continue with our discussion and introduction of inferential statistics. Recall that as we analyze a data set, we generally want to begin by describing it (central tendency, measures of variability), and we also want to plot the data. To begin our introduction to correlation and regression, first we describe how to produce graphs to help show **linear association** or in some cases, cause and effect — the latter perhaps the primary reason for using regression.

Graphical representation

The previous statistical procedures we have examined have used one or more **categorical** or **qualitative** variables (Chapter 3). For example,

1. Chi-Square Analyses: variables are all categorical, including the response variable (Chapter 9).
2. T-tests: one categorical (Factor) variable and one (Dependent, Outcome, Response) variable that was continuous or interval scale (Chapter 8.5, 10).
3. ANOVA Analyses: one or more variables are categorical (Factors, the independent variables) and one (Dependent, Outcome, Response) variable that was continuous or interval scale (Chapter 12, 14).

The convention in graphing ANOVA (or Chi-Square) is to use the Factor or Independent variables as the X-axis and to have the dependent variable (Response) as the Y-axis. We called these **bar charts** (Chapter 4.1).

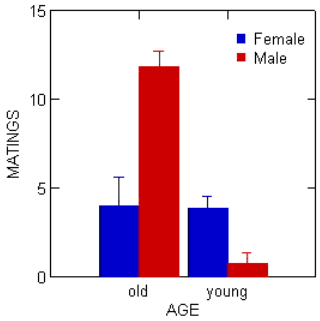


Figure 16.1: Bar chart with error bars.

Box plots (Chapter 4.3) are also useful, and perhaps the preferred choice to display this type of comparison (one involving groups) (Fig. 16.2).

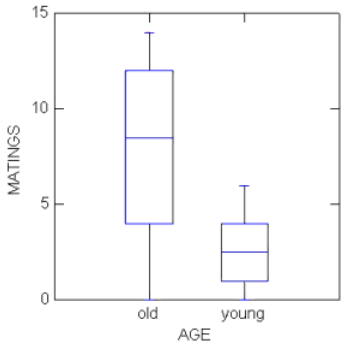


Figure 16.2: Box plots.

In correlation (and regression) analyses we will have two or more continuous or interval scale variables. To show relationships among continuous variables, a **scatter plot,** also called an X-Y plot, works well (Chapter 4.5).

In correlation, no causation is implied, so either variable can be placed on the X-axis. The convention of graphing in regression is to place the independent variable as the **X-axis** and the dependent variable as the **Y-axis** (Fig. 16.3). Another consideration: if one

variable is considered fixed and the other random, then the fixed variable would be assigned to the horizontal axis.

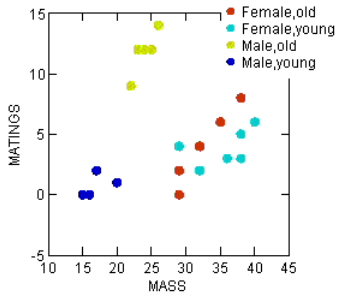


Figure 16.3: Scatterplot with groups.

To produce a scatterplot (also called an X-Y plot) in `Rcmdr` , select **Graph** → **Plot** → and select the Y and X variables. Use a combination of Options, Frame, and Edit Attributes selections to modify the default graph.

16.1: Product-moment correlation

Introduction

A correlation is used to describe the direction $\pm$ the magnitude of **linear association** between two variables. There are many types of correlations; some are based on ranks, but the one most commonly used is the product-moment correlation (r). The Pearson **product-moment correlation** is used to describe association between continuous, ratio-scale data, where "Pearson" is in honor of Karl Pearson (b. 1857 – d. 1936).

There are many other correlations, including Spearman's and Kendall's tau (τ) (Chapter 16.4) and ICC, the **intraclass correlation** (Chapter 12.3 and Chapter 16.4).

The product moment correlation is appropriate for variables of the same kind — for example, two measures of size, like the correlation between body weight and brain weight.

Spearman's and Kendall's tau correlation are nonparametric and would be alternatives to the product moment correlation. The intraclass correlation, or ICC, is a parametric estimate suitable for repeat measures of the same variable.

The correlation coefficient

$$r_{XY} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{(n-1)s_X s_Y}$$

The numerator is the sum of products and it quantifies how the deviates from the X and Y means covary, or change together. The numerator is known as a "covariance."

$$COV = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

The denominator includes the standard deviations of X and Y; thus, the correlation coefficient is the standardized covariance.

The product moment correlation, r, is an estimate of the population correlation, ρ (pronounced rho), the true relationship between the two variables.

$$\rho_{XY} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

where $COV(X,Y)$ refers to the covariance between X and Y.

Effect size

Estimates for r range from -1 to $+1$; the correlation coefficient has no units. A value of 0 describes the case of no statistical correlation, i.e., no linear association between the two variables. Usually, this is taken as the null hypothesis for correlation — "No correlation between two variables," with the alternative hypothesis (2-tailed) — "There is a correlation between two variables."

Like effect size, we can report the strength of correlation between two variables. Consider the magnitude and not the direction (\pm\). Like Cohen's effect size:

Absolute value	Magnitude of association
0.10	small, weak
0.30	moderate
< 0.50	strong, large

Note that one should not interpret a "strong, large" correlation as evidence that the association is necessarily real. See Chapter 16.2 for more on **spurious correlations**.

Standard error of the correlation

An approximate standard error for r can be obtained using this simple formula:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

This standard error can be used for significance testing with the t-test. See below.

Confidence interval

Like all situations in which an estimate is made, you should report the confidence interval for r. The standard error approximation is appropriate when the null hypothesis is $r = 0$, because the joint distribution is approximately normal. However, as the estimate approaches the limits of the closed interval $[-1, 1]$, the distribution becomes increasingly skewed.

The approximate confidence interval for the correlation is based on **Fisher's z-transformation**. We use this transformation to stabilize the variance over the range of possible values of the correlation and, therefore, better meet the assumptions of parametric tests based on the normal distribution.

The transform is given by the equation

$$z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

where ln is the natural logarithm. In the R language we get the natural log by `log(x)`, where x is a variable we wish to transform.

Equivalently, z can be rewritten as

$$z = arctanh(r)$$

using the **inverse hyperbolic tangent** function. In R language this function is called by `atanh(r)` at the R prompt.

The standard error for z is about

$$\sigma_z = \sqrt{\frac{1}{n-3}}$$

We take z to be the estimate of the population zeta, ζ. We take the sampling distribution of z to be approximately normal, and thus we may then use the normal table to generate the 95% confidence interval for zeta.

$$z - 1.96_z < \zeta < z + 1.96_z$$

Why 1.96? We want 95% confidence interval, so that at Type I $\alpha = 0.05$; we want the two tails of the Normal distribution (see Appendix 20.1), so we divide the 0.05 value by 2 to get 0.025. Thus $+0.025$ is $+1.96$ and -1.96 corresponds to -0.025.

Significance testing

Significance testing of correlations is straightforward, with the noted caveat about the need to transform in cases where the estimate is close to ± 1. For the typical test of null hypothesis, the correlation, r, is equal to 0, and the t distribution can be used (i.e., it's a t-test).

$$t = \frac{r-0}{s_r}$$

which has degrees of freedom $DF = n - 2$.

Use the t-table critical values to test the null hypothesis involving product moment correlation (e.g., Appendix 4; for Spearman rank correlation r_s see Table G, p. 686 in Whitlock & Schluter).

Alternatively (and preferred), we'll just use R and Rcmdr's facilities without explanation; the t distribution works OK as long as the correlations are not close to ± 1, in which case other things need to be done — and this is also true if you want to calculate a confidence interval for the correlation.

You are sufficiently skilled at this point to evaluate whether a correlation is statistically significantly different from zero — just check out whether the associated p-value is less than or greater than alpha (usually set at 5%). A test of whether or not the correlation, r_1, is equal to some value, r_2, other than zero is also possible. For an approximate test, replace zero in the above test statistic calculation with the value for r_2, and calculate the standard error of the difference. Note that use of the t-test for

significance testing of the correlation is an approximate test — if the correlations are small in magnitude using the Fisher's z transformation approach will be less biased, where the test statistic z now is

$$z = \frac{z_{r_1} - z_{r_2}}{\sigma_{z_1 - z_2}}$$

and standard error of the difference is

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} - \frac{1}{n_2 - 3}}$$

and look up the critical value of z from the normal table.

R code

To calculate correlations in R and Rcmdr, have ratio-scale data ready in the columns of a R and Rcmdr data frame. We'll introduce the commands with an example data set from my genetics laboratory course.

Question. What is the estimate of the product moment correlation between *Drosophila* fly wing length and area?

Data (thanks to some of my genetics students!)

```
Area <- c(0.446, 0.876, 0.390, 0.510, 0.736, 0.453, 0.882, 0.394, 0.503, 0.535, 0.441
Length <- c(1.524, 2.202, 1.520, 1.620, 1.710, 1.551, 2.228, 1.460, 1.659, 1.719, 1.5
```

Create your data frame, e.g.,

```
FlyWings <- data.frame(Area, Length)
```

And here's the scatterplot. We can clearly see that Wing Length and Wing Area are positively correlated, with one outlier (Fig. 16.1.1).

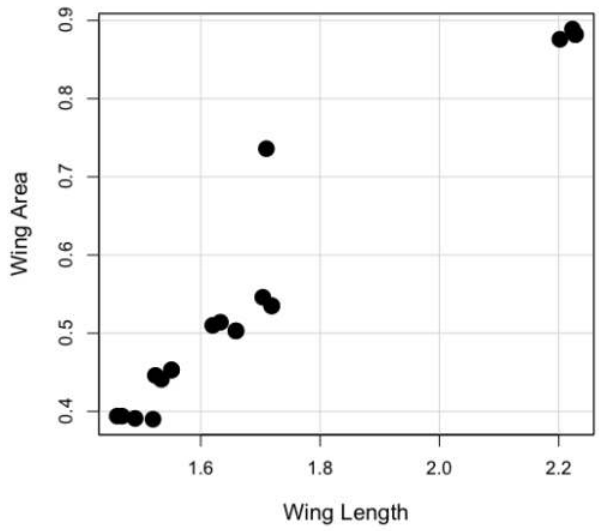


Figure 16.1.1: Scatterplot of *Drosophila* wing area by wing length.

The R command for correlation is simply `cor(x,y)` . This gives the "pearson" product moment correlation, the default. To specify other correlations, use method = "kendall", or method = "spearman" (See Chapter 16.4).

Question. What are the Pearson, Spearman, and Kendall's tau estimates for the correlation between fly Wing Length and Wing Area?

At the R prompt, type

```
cor(Length,Area)
[1] 0.9693334
cor(Length,Area, method="kendall")
[1] 0.8248008
cor(Length,Area, method="spearman")
[1] 0.9558658
```

Note that we entered `Length` first. On your own, confirm that the order of entry does not change the correlation estimate. To both estimate test the significance of the correlation between Wing Area and Wing Length, at the R prompt type

```
cor.test(Area, Length, alternative="two.sided", method="pearson")
```

R returns with

```
Pearson's product-moment correlation

data: Area and Length

t = 16.735, df = 18, p-value = 2.038e-12>
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9225336 0.9880360
sample estimates:
   cor
0.9693334
```

Alternatively, to calculate and test the correlation, use R Commander, **Rcmdr: Statistics → Summaries→ Correlation test**

Note:

R's `cor.test` uses Fisher's z transformation; note if we instead use the approximate calculation instead how poor the approximation works in this example. The estimated correlation was 0.97, thus the approximate standard error was 0.058. The confidence interval (t-distribution, $\alpha = 0.05/2$ and 18 degrees of freedom) was between 0.848 and 1.091, which is greater than the z transform result and returns an out-of-bounds upper limit.

Alternative packages to base R provide more flexibility and access to additional approaches to significance testing of correlations (Goertzen and Cribbie 2010). For example, `z_cor_test()` from the `TOSTER` package.

```
z_cor_test(Area, Length)

Pearson's product-moment correlation
data:  Area and Length
z = 8.5808, N = 20, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9225336 0.9880360
sample estimates:
      cor
0.9693334
```

To confirm, check the critical value for $z = 8.5808$, two-tailed, with

```
> 2*pnorm(c(8.5808), mean=0, sd=1, lower.tail=FALSE)
[1] 9.421557e-18
```

Note the difference is that Fisher's z is used for hypothesis testing; `cor.test` and `z_cor_test` return the same confidence intervals.

We could also use bootstrap resampling (see Chapter 19.2),

```
boot_cor_test(Area, Length)

Bootstrapped Pearson's product-moment correlation

data: Area and Length
N = 20, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8854133 0.9984744
sample estimates:
cor
0.9693334
```

The z-transform confidence interval would be preferred over the bootstrap confidence interval because it is narrower.

Assumptions of the product-moment correlation

Interestingly enough, there are no assumptions for estimating a statistic. You can always calculate an estimate, although of course, this does not mean that you have selected the best calculation to describe the phenomenon in question; it just means that assumptions are not applicable for estimation. Whether it is the sample mean or the correlation, it is important to appreciate that, look, you can always calculate it, even if it is not appropriate!

Statistical assumptions and those technical hypotheses we evaluate apply to statistical inference — being able to correctly interpret a test of statistical significance for a correlation estimate depends on how well assumptions are met. The most important assumption for a null hypothesis test of correlation is that samples were obtained from a "bivariate normal distribution." It is generally sufficient to just test normality of the variables one at a time (univariate normality), but the student should be aware that testing the bivariate normality assumption can be done directly (e.g., Doornick and Hansen 2008).

Testing two independent correlations

Extending from a null hypothesis of the correlation is equal to zero to the correlation equals a particular value should not be a stretch for you. For example, since we use the t-test to evaluate the null hypothesis that the correlation is equal to zero, you should be able to make the connection that, like the two sample t-test, we can extend the test of correlation to any value. However, using the t-test without considering the need to stabilize the variance.

When two correlations come from independent samples, we can test whether or not the two correlations are equal. Rather than use the t-test, however, we use a modification of Fisher's Z transformation. Calculate z for each correlation separately, then use the following equation to obtain Z. We then look up Z from our table of standard normal distribution (Appendix A,2, or better — use the normal distribution functions in `Rcmdr`) and we can obtain the p-value of the test of the hypothesis that the two correlations are equal.

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Example. Two independent correlations are $r_1 = 0.2$ and $r_2 = 0.34$. Sample size for group 1 was 14 and for group 2 was 21. Test the hypothesis that the two correlations are equal.

Using R as a calculator, here's what we might write in the R script window and the resulting output. It doesn't matter which correlation we set as r1 or r2, so I prefer to calculate the absolute value of Z and then get the probability from the normal table for values greater or equal to $|Z|$ (i.e., the upper tail).

```
z1 = atanh(0.2)
z2 = atanh(0.34)
n1 = 14
n2 = 21
Z = abs((z1-z2)/sqrt((1/(n1-3))+(1/(n2-3))))
Z = 0.3954983
```

From the normal distribution table we get a p-value of 0.3462 for the upper tail. Because this p-value is not less than our typical Type I error rate of 0.05, we conclude that the two correlations are not in fact significantly different.

Rcmdr: Distributions → Continuous distributions → Normal distribution → Normal probabilities…

```
pnorm(c(0.3954983), mean=0, sd=1, lower.tail=FALSE)
```

R returns

```
[1] 0.3462376
```

To make this two-tailed, of course all we have to do is multiple the one-tailed p-value by two; in this case the two-tailed p-value = 0.69247.

Write a function in R

There's nothing wrong with running the calculations as written, but R allows users to write their own functions. Here's one possible function we could write to test two independent correlations. Write the R function in the script window.

```
test2Corr = function(r1,r2,n1,n2) {
z1=atanh(r1); z2=atanh(r2)
Z = abs((z1-z2)/sqrt((1/(n1-3))+(1/(n2-3))))
pnorm(c(Z), mean=0, sd=1, lower.tail=FALSE)
}
```

After submitting the function, we then invoke the function by typing at the R prompt

```
p = test2Corr(0.2,0.34,14,21); p
```

Again, R returns the one-tailed p-value

```
[1] 0.3462376
```

Unsurprisingly, these simple functions are often available in an R package. In this case, the psych package provides a function called `r.test()` which will accommodate the test of the equality hypothesis of two independent correlations. Assuming that the psych package has been installed, at the R prompt we type

```
require(psych)
r.test(14,.2,.34,n2=21,twotailed=TRUE)
```

And R returns

```
Correlation tests
Call:r.test(n = 14, r12 = 0.20, r34 = 0.34, n2 = 21, twotailed = TRUE)
Test of difference between two independent correlations
z value 0.4 with probability 0.69
```

Questions

1. True or False. It is relatively easy to move from the estimation of one correlation between two continuous variables, to the estimation of multiple pairwise ("2 at a time") correlations among many variables. For k = the number of variables, there are $k(k-1)/2$ unique correlations. However, one should be concerned about the multiple comparisons problem as introduced in ANOVA when one tests for the statistical significance of many correlations.
2. True or False. Generally, the null hypothesis of a test of a correlation is $H_O : r = 0$, although in practice, one could test a null of r = any value.
3. Return to the fly wing example. What was the estimate of the value of the product moment correlation? The Spearman Rank correlation? The Kendall's tau?
4. OK, you have three correlation estimates for test of the same null hypothesis, i.e., correlation between Length and Area is zero. Which estimate is the best estimate?
5. Apply the Fisher z transformation to the estimated correlation, what did you get?
6. For the fly wing example, what were the degrees of freedom?
7. For the fly wing example, calculate the approximate standard error of the product moment correlation.
8. Return one last time to the fly wing example. What was the value of the lower limit of the 95% confidence interval for the estimate of the product moment correlation? And the value of the upper limit?
9. Assume that another group of students $(n = 15)$ made measurements on fly wings and the correlation was 0.86. Is the difference between the two correlations for the two groups of students equal? Obtain the probability using the Z calculation and R (Chapter 6.7) or the normal table.

16.2: Causation and partial correlation

Introduction

Science driven by statistical inference and model building is largely motivated by the the drive to identify pathways of **cause and effect** linking events and phenomena observed all around us. (We first defined cause and effect in Chapter 2.4) The history of philosophy, from the works of Ancient Greece, China, Middle East and so on is rich in the language of cause and effect. From these traditions we have a number of ways to think of cause and effect, but for us it will be enough to review the logical distinction among three kinds of cause-effect associations:

- Necessary cause
- Sufficient cause
- Contributory cause

Here's how the logic works. If ***A*** is a **necessary cause** of ***B***, then the mere fact that ***B*** is present implies that ***A*** must also be present. Note, however, that the presence of ***A*** does not imply that ***B*** will occur. If ***A*** is a **sufficient cause** of ***B***, then the presence of ***A*** necessarily implies the presence of ***B***. However, another cause ***C*** may alternatively cause ***B***. Enter the contributory or related cause: A cause may be **contributory** if the presumed cause ***A*** (1) occurs before the effect ***B***, and (2) changing ***A*** also changes ***B***. Note that a contributory cause does not need to be necessary nor must it be sufficient; contributory causes play a role in cause and effect.

Thus, following this long tradition of thinking about causality, we have the mantra "Correlation does not imply causation." The exact phrase was written as early as the late 1800s, when it was emphasized by Karl Pearson, who invented the correlation statistic. This well-worn slogan deserves to be on T-shirts and bumper stickers*, and perhaps to be viewed as the single most important concept you can take from a course in philosophy/statistics. But in practice, we will always be tempted to stray from this guidance. The developments in genome-wide-association studies, or GWAS, are designed to look for correlations, as evidenced by statistical linkage analysis, between variation at one DNA base pair and presence/absence of disease or condition in humans and animal models. These are costly studies to do and in the end, the results are just that, evidence of associations (correlations), not proof of genetic cause and effect. We are less likely to be swayed by a correlation that is weak, but what about correlations that are large, even close to one? Is not the implication of high, statistically significant correlation evidence of causation? No, necessary, but not sufficient.

Note:

A helpful review on causation in epidemiology is available from Parascandola and Weed (2001); see also Kleinberg and Hripcsak (2011). For more on "correlation does not imply causation", try the Wikipedia entry. Obviously, researchers who engage in genome wide association studies are aware of these issues: see for example discussion by Hu et al (2018) on **causal inference** and GWAS.

Causal inference (Pearl 2009; Pearl and Mackenzie 2018), in brief, employs a model to explain the association between dependent and multiple, likely interrelated candidate causal variable, which is then subject to testing — is the model stable when the predictor variables are manipulated, when additional connections are considered (e.g., predictor variable 1 covaries with one or more other predictor variables in the model). Wright's path analysis, now included as one approach to **Structural Equation Modeling**, is used to relate equations (models) of variation in observed variables attributed to direct and indirect effects from predictor variables.

* And yes, a quick Google search reveals lots of bumper stickers and T-shirts available with the causation $\neq$ sentiment.

Spurious correlations

Correlation estimates should be viewed as hypotheses in the scientific sense of the meaning of hypotheses for putative cause-effect pairings. To drive the point home, explore the web site "Spurious Correlations" at https://www.tylervigen.com/spurious-correlations , which allows you to generate X-Y plots and estimate correlations among many different variables. Some of my favorite correlations from "Spurious Correlations" include (Table 16.2.1):

Table 16.2.1. Spurious correlations, https://www.tylervigen.com/spurious-correlations

First variable	Second variable	Correlation
Divorce rate in Maine, USA	Per capita USA consumption of margarine	+0.993

First variable	Second variable	Correlation
Honey producing bee colonies USA	Juvenile arrests for marijuana possession	-0.933
Per capita USA consumption of mozzarella cheese	Civil engineering PhD awarded USA	+0.959
Total number of ABA lawyers USA	Cost of red delicious apples	+0.879

These are some pretty strong correlations (cf. effect size discussion, Ch. 11.4), about as close to $+1$ as you can get. But really, do you think the amount of cheese that is consumed in the USA has anything to do with the number of PhD degrees awarded in engineering or that apple prices are largely set by the number of lawyers in the USA? Cause and effect implies there must also be some plausible mechanism, not just a strong correlation.

But that does NOT ALSO mean that a high correlation is meaningless. The primary reason a correlation cannot tell about causation is because of the problem (potentially) of an UNMEASURED variable (a **confounding variable**) being the real driving force (Fig. 16.2.1).

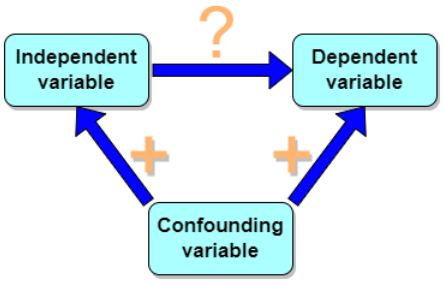


Figure 16.2.1: Unmeasured confounding variables influence association between independent and dependent variables, the characters or traits we are interested in.

Here's a plot of running times for the fastest men and women runners for the 100-meter sprint, since the 1920s. The data are collated for you and presented at end of this page (scroll or click here).

Here's a scatterplot (Fig. 16.2.2).

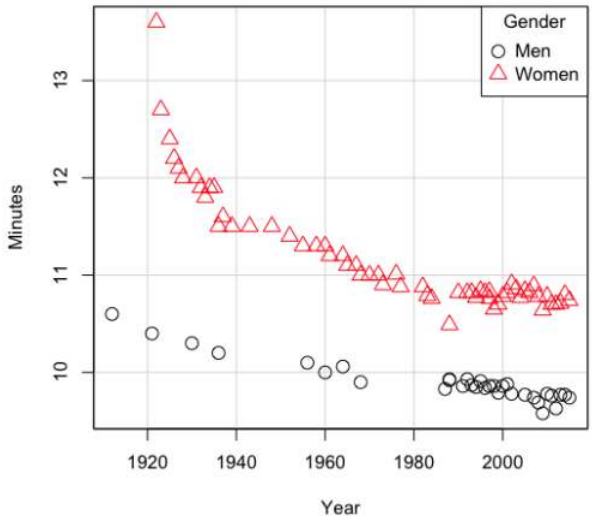


Figure 16.2.2: Running times over 100 meters of top athletes since the 1920s.

There's clearly a negative correlation between years and running times. Is the rate of improvement in running times the same for men and women? Is the improvement linear? What, if any, are the possible confounding variables? Height? Weight? Biomechanical differences? Society? Training? Genetics? … Performance enhancing drugs…?

If we measure potential confounding factors, we may be able to determine the strength of correlation between two variables that share variation with a third variable.

The partial correlation

There are several ways to work this problem. The partial correlation is a useful way to handle this problem, i.e., where a measured third variable is positively correlated with the two variable you are interested in.

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2} \cdot \sqrt{1 - r_{23}^2}}$$

Without formal mathematical proof presented, \(r_{12.3\) is the correlation between variables 1 and 2 INDEPENDENT of any covariation with variable 3.

For our running data set, we have the correlation between women's time for 100 m over 9 decades, $(r_{13} = -0.876)$, between men's time for 100 m over 9 decades $(r_{23} = -0.952)$, and finally, the correlation we're interested in, whether men's and women's times are correlated $(r_{12} = +0.71)$. When we use the partial correlation, however, I get $r_{12.3} = -0.819\ldots$ much less than 0 and significantly different from zero. In other words, men's and women's times are not positively correlated independent of the correlation both share with the passage of time (decades)! The interpretation is that men are getting faster at a rate faster than women.

In conclusion, keep your head about you when you are doing analyses. You may not have the skills or knowledge to handle some problems (partial correlation), but you can think simply — why are two variables correlated? One causes the other to increase (or decrease) OR the two are both correlated with another variable.

Testing the partial correlation

Like our simple correlation, the partial correlation may be tested by a t-test, although modified to account for the number of pairwise correlations (Wetzels and Wagenmakers 2012). The equation for the t test statistic is now

$$t = r_{12.3}\sqrt{\frac{n - 2 - k}{1 - r_{12.3}^2}}$$

with k equal to the number of pairwise correlations and $n - 2 - k$ degrees of freedom.

Examples

Lead exposure and low birth weight. The data set is numbers of low birth weight births (< 2,500 g regardless of gestational age) and numbers of children with high levels of lead (10 or more micrograms of lead in a deciliter of blood) measured from their blood. Data used for 42 cities and towns of Rhode Island, United States of America (data at end of this page, scroll or click here to access the data).

A scatterplot of number of children with high lead is shown below (Fig. $16.2.3$).

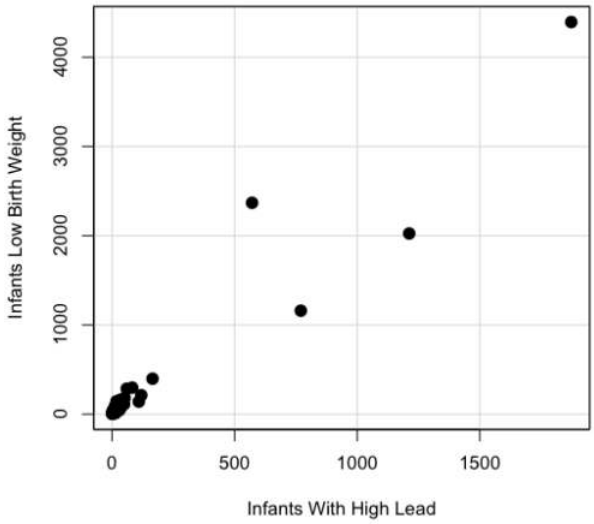


Figure 16.2.3: Scatterplot birth weight by lead exposure.

The product moment correlation was $r = 0.961$, $t = 21.862$, $df = 40$, $p < 2.2 \times 10^{-16}$. So, at first blush looking at the scatterplot and the correlation coefficient, we conclude that there is a significant relationship between lead and low birth weight, right?

However, by the description of the data you should note that counts were reported, not rates (e.g., per 100,000 people). Clearly, population size varies among the cities and towns of Rhode Island. West Greenwich had 5085 people whereas Providence had 173,618. We should suspect that there is also a positive correlation between number of children born with low birth weight and numbers of children with high levels of lead. Indeed there are.

Correlation between Low Birth Weight and Population, $r = 0.982$

Correlation between High Lead levels and Population, $r = 0.891$

The question becomes, after removing the covariation with population size is there a linear association between high lead and low birth weight? One option is to calculate the partial correlation. To get partial correlations in `Rcmdr` , select

Statistics → Summaries → Correlation matrix

then select "partial" and select all three variables (Ctrl key) (Fig. 16.2.4)

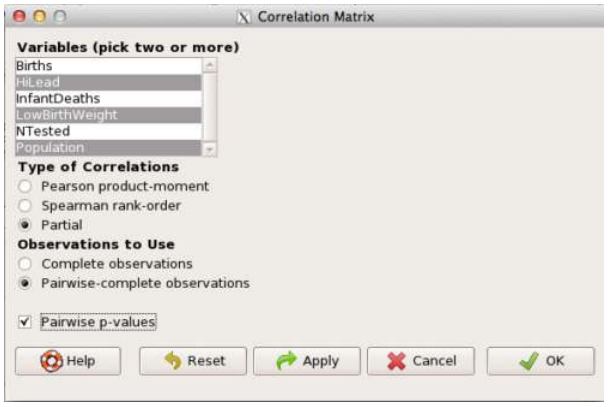


Figure 16.2.4: Screenshot of Rcmdr partial correlation menu.

Results are shown below.

```
partial.cor(leadBirthWeight[,c("HiLead","LowBirthWeight","Population")], tests=TRUE,

Partial correlations:
                HiLead LowBirthWeight Population
```

```
HiLead          0.00000        0.99181   -0.97804
LowBirthWeight  0.99181        0.00000    0.99616
Population     -0.97804        0.99616    0.00000
```

Thus, after removing the covariation we conclude there is indeed a strong correlation between lead and low birth weights.

Note:

A little bit of verbiage about correlation tables (matrices). Note that the matrix is symmetric and the information is repeated. I highlighted the diagonal in green. The upper triangle (red) is identical to the lower triangle (blue). When you publish such matrices, don't publish both the upper and lower triangles; it's also not necessary to publish the on-diagonal numbers, which are generally not of interest. Thus, the publishable matrix would be

```
Partial correlations:
                LowBirthWeight Population
HiLead                 0.99181   -0.97804
LowBirthWeight                    0.99616
```

Another example

Do Democrats prefer cats? The question I was interested in, Do liberals really prefer cats?, was inspired by a *Time* magazine 18 February 2014 article. I collated data on a separate but related question: Do states with more registered Democrats have more cat owners? The data set was compiled from three sources: 2010 USA Census, a 2011 Gallup poll about religious preferences, and from a data book on survey results of USA pet owners (data at end of this page, scroll or click here to access the data).

Note:

This type of data set involves questions about groups, not individuals. We have access to aggregate statistics for groups (city, county, state, region), but not individuals. Thus, our conclusions are about groups and cannot be used to predict individual behavior, e.g., knowing a person votes Green Party does not mean they necessarily share their home with a cat). See ecological fallacy.

This data set also demonstrates use of **transformations of the data** to improve fit of the data to statistical assumptions (**normality**, **homoscedacity**).

The variables, and their definitions, were:

ASDEMS = DEMOCRATS. Democrat advantage: the difference in registered Democrats compared to registered Republicans as a percentage; to improve the distribution qualities the arcsine transform was applied..

ASRELIG = RELIGION. Percent Religous from a Gallup poll who reported that Religion was "Very Important" to them. Also arcsine-transformed to improve normality and **homoescedasticity** (there you go, throwing $3 words around 🤠).

LGCAT = Number of pet cats, $\log_{10}$-transformed, estimated for USA states by survey, except Alaska and Hawaii (not included in the survey by the American Veterinary Association).

LGDOG = Estimated number of pet dogs, $\log_{10}$-transformed for states, except Alaska and Hawaii (not included in the survey by the American Veterinary Association).

LGIPC = Per capita income, $\log_{10}$-transformed.

LGPOP = Population size of each state, $\log_{10}$ transformed.

As always, begin with data exploration. All of the variables were right-skewed, so I applied data transformation functions as appropriate: $\log_{10}$ for the quantitative data and arcsine transform for the frequency variables. Because Democrat Advantage and Percent Religious variables were in percentages, the values were first divided by 100 to make frequencies, then the R function `asin()` was applied. All analyses were conducted on the transformed data, therefore conclusions apply to the transformed data.

To relate the results to the original scales, back transformations would need to be run on any predictions. Back transformation for $\log_{10}$ would be power of ten; for the arcsine-transform the inverse of the arcsine would be used.

A scatter plot matrix (`KMggplo2`) plus histograms of the variables along the diagonals shows the results of the transforms and hints at the associations among the variables. A graphic like this one is called a trellis plot; a layout of smaller plots in a grid with the same (preferred) or at least similar axes. **Trellis plots** (Fig. 16.2.5) are useful for finding the structure and patterns in complex data. Scanning across a row shows relationships between one variable with all of the others. For example, the first row Y-axis is for the ASDEMS variable; from left to right along the row we have, after the histogram, what look to be weak associations between ASDEMS and ASRELIG, LGCAT, LGDOG, and LGDOG.

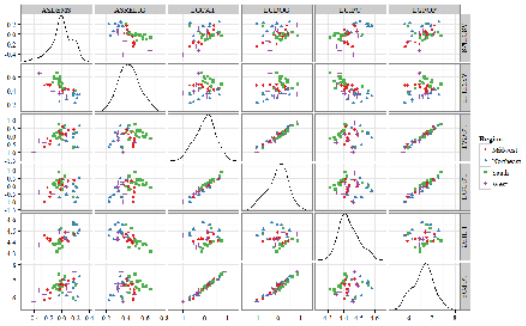


Figure 16.2.5: Trellis plot, correlations among variables.

A matrix of partial correlations was produced from the `Rcmdr` correlation call. Thus, to pick just one partial correlation, the association between DEMOCRATS and RELIGION (reported as "very important") is negative $(r = -0.45)$ and from the second matrix we retrieve the approximate p-value, unadjusted for the multiple comparisons problem, of $p = 0.0024$. We quickly move past this matrix to the adjusted p-values and confirm that this particular correlation is statistically significant even after correcting for multiple comparisons. Thus, there is a moderately strong negative correlation between those who reported that religion was very important to them and the difference between registered Democrats and Republicans in the 48 states. Because it is a partial correlation, we can conclude that this correlation is independent of all of the other included variables.

And what about our original question: Do Democrats prefer cats over dogs? The partial correlation after adjusting for all of the other correlated variables is small $(r = 0.05)$ and not statistically different from zero (p-value greater than 5%).

Are there any interesting associations involving pet ownership in this data set? See if you can find it (hint: the correlation you are looking for is also in red).

```
Partial correlations:
          ASDEMS  ASRELIG LGCAT   LGDOG   LGIPC   LGPOP
ASDEMS    0.0000 -0.4460  0.0487  0.0605  0.1231 -0.0044
ASRELIG -0.4460  0.0000 -0.2291 -0.0132 -0.4685  0.2659
LGCAT     0.0487 -0.2291  0.0000  0.2225 -0.1451  0.6348
LGDOG     0.0605 -0.0132  0.2225  0.0000 -0.6299  0.5953
LGIPC     0.1231 -0.4685 -0.1451 -0.6299  0.0000  0.6270
LGPOP    -0.0044  0.2659  0.6348  0.5953  0.6270  0.0000
```

Raw P-values, Pairwise two-sided p-values:

```
          ASDEMS ASRELIG  LGCAT   LGDOG   LGIPC   LGPOP
ASDEMS           0.0024  0.7534  0.6965  0.4259  0.9772
ASRELIG  0.0024          0.1347  0.9325  0.0013  0.0810
LGCAT    0.7534  0.1347          0.1465  0.3473  <.0001
LGDOG    0.6965  0.9325  0.1465          <.0001  <.0001
LGIPC    0.4259  0.0013  0.3473  <.0001          <.0001
LGPOP    0.9772  0.0810  <.0001  <.0001  <.0001
```

Adjusted P-values, Holm's method (Benjamini and Hochberg 1995)

```
          ASDEMS ASRELIG  LGCAT   LGDOG   LGIPC   LGPOP
ASDEMS            0.0241 1.0000  1.0000  1.0000  1.0000
ASRELIG 0.0241           1.0000  1.0000  0.0147  0.7293
LGCAT   1.0000  1.0000           1.0000  1.0000  <.0001
LGDOG   1.0000  1.0000  1.0000           <.0001  0.0002
LGIPC   1.0000  0.0147  1.0000  <.0001           <.0001
LGPOP   1.0000  0.7293  <.0001  0.0002  <.0001
```

A graph (Fig. 16.2.6) to summarize the partial correlations: green lines indicate positive correlation, red lines show negative correlations. Strength of association is indicated by the line thickness, with thicker lines corresponding to greater correlation.

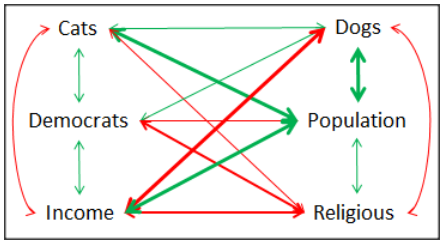


Figure 16.2.6: Causal paths among variables.

As you can see, partial correlation analysis is good for a few variables, but as the numbers increase it is difficult to make heads or tails out of the analysis. Better methods for working with these highly correlated data in what we call multivariate data analysis, for example **Structural Equation Modeling** or **Path Analysis**.

Questions

1. True of False. We know that correlations should not be interpreted as "cause and effect." However, it is safe to assume that a correlation very close to the limits (r = 1 or r = -1) is likely to mean that one of the variables causes the other to vary.
2. Spurious correlations can be challenging to recognize, and, sometimes, they become part of a challenge to medicine to explain away. A classic spurious correlation is the correlation between rates of MMR vaccination and autism prevalence. Here's a table of numbers for you.

Table 16.2.2. Autism rates and additional "causal" variables.

Year	Herb Supplement Revenue, Millions	Fertility rate per 1000 births, women aged 35 and over	MMR per 100K children age 0-5	UFC revenue, millions	Autism prevalence per 1000
2000	4225	47.7	179		6.7
2001	4361	48.6	183	4.5	
2002	4275	49.9	190	8.7	6.6
2003	4146	52.6	196	7.5	
2004	4288	54.5	199	14.3	8
2005	4378	55.5	197	48.3	
2006	4558	56.9	198	180	9
2007	4756	57.6	204	226	
2008	4800	56.7	202	275	11.3
2009	5037	56.1	201	336	

2010	5049	56.1	209	441	14.4
2011	5302	57.5	212	437	
2012	5593	58.7	216	446	14.5
2013	6033	59.7	220	516	
2014	6441	61.6	224	450	16.8
2015	6922	62.8	222	609	
2016	7452	64.1	219	666	18.5
2017	8085	63.9	213	735	
2018		65.3	220	800	25

3. Make scatterplots of autism prevalence vs
    - Herb supplement revenue
    - Fertility rate
    - MMR vaccination
    - UFC revenue
4. Calculate and test correlations between autism prevalence vs
    - Herb supplement revenue
    - Fertility rate
    - MMR vaccination
    - UFC revenue
5. Interpret the correlations — is there any clear case for autism vs MMR?
6. What additional information is missing from Table 2? Add that missing variable and calculate partial correlations for autism prevalence vs
    - Herb supplement revenue
    - Fertility rate
    - MMR vaccination
    - UFC revenue
7. Do a little research: What are some reasons for increase in autism prevalence? What is the consensus view about MMR vaccine and risk of autism?

Data used in this page, 100 meter running times since 1900.

Year	Men	Women
1912	10.6	
1913		
1914		
1915		
1916		
1917		
1918		
1919		
1920		
1921	10.4	

1922		13.6
1923		12.7
1924		
1925		12.4
1926		12.2
1927		12.1
1928		12
1929		
1930	10.3	
1931		12
1932		11.9
1933		11.8
1934		11.9
1935		11.9
1936	10.2	11.5
1937		11.6
1938		
1939		11.5
1940		
1941		
1942		
1943		11.5
1944		
1945		
1946		
1947		
1948		11.5
1949		
1950		
1951		
1952		11.4
1953		
1954		
1955		11.3
1956	10.1	

1957		
1958		11.3
1959		
1960	10	11.3
1961		11.2
1962		
1963		
1964	10.06	11.2
1965		11.1
1966		
1967		11.1
1968	9.9	11
1969		
1970		11
1972	10.07	11
1973	10.15	10.9
1976	10.06	11.01
1977	9.98	10.88
1978	10.07	10.94
1979	10.01	10.97
1980	10.02	10.93
1981	10	10.9
1982	10	10.88
1983	9.93	10.79
1984	9.96	10.76
1987	9.83	10.86
1988	9.92	10.49
1989	9.94	10.78
1990	9.96	10.82
1991	9.86	10.79
1992	9.93	10.82
1993	9.87	10.82
1994	9.85	10.77
1995	9.91	10.84
1996	9.84	10.82

1997	9.86	10.76
1998	9.86	10.65
1999	9.79	10.7
2000	9.86	10.78
2001	9.88	10.82
2002	9.78	10.91
2003	9.93	10.86
2004	9.85	10.77
2005	9.77	10.84
2006	9.77	10.82
2007	9.74	10.89
2008	9.69	10.78
2009	9.58	10.64
2010	9.78	10.78
2011	9.76	10.7
2012	9.63	10.7
2013	9.77	10.71
2014	9.77	10.8
2015	9.74	10.74
2016	9.8	10.7
2017	9.82	10.71
2018	9.79	10.85
2019	9.76	10.71
2020	9.86	10.85

Data used in this page, birth weight by lead exposure

CityTown	Core	Population	NTested	HiLead	Births	LowBirthWeight	InfantDeaths
Barrington	n	16819	237	13	785	54	1
Bristol	n	22649	308	24	1180	77	5
Burrillville	n	15796	177	29	824	44	8
Central Falls	y	18928	416	109	1641	141	11
Charlestown	n	7859	93	7	408	22	1
Coventry	n	33668	387	20	1946	111	7
Cranston	n	79269	891	82	4203	298	20
Cumberland	n	31840	381	16	1669	98	8
East Greenwich	n	12948	158	3	598	41	3

CityTown	Core	Population	NTested	HiLead	Births	LowBirthWeight	InfantDeaths
East Providence	n	48688	583	51	2688	183	11
Exeter	n	6045	73	2	362	6	1
Foster	n	4274	55	1	208	9	0
Glocester	n	9948	80	3	508	32	5
Hopkintown	n	7836	82	5	484	34	3
Jamestown	n	5622	51	14	215	13	0
Johnston	n	28195	333	15	1582	102	6
Lincoln	n	20898	238	20	962	52	4
Little Compton	n	3593	48	3	134	7	0
Middletown	n	17334	204	12	1147	52	7
Narragansett	n	16361	173	10	728	42	3
Newport	y	26475	356	49	1713	113	7
New Shoreham	n	1010	11	0	69	4	1
North Kingstown	n	26326	378	20	1486	76	7
North Providence	n	32411	311	18	1679	145	13
North Smithfield	n	10618	106	5	472	37	3
Pawtucket	y	72958	1125	165	5086	398	36
Portsmouth	n	17149	206	9	940	41	6
Providence	y	173618	3082	770	13439	1160	128
Richmond	n	7222	102	6	480	19	2
Scituate	n	10324	133	6	508	39	2
Smithfield	n	20613	211	5	865	40	4
South Kingstown	n	27921	379	35	1330	72	10
Tiverton	n	15260	174	14	516	29	3
Warren	n	11360	134	17	604	42	1
Warwick	n	85808	973	60	4671	286	26
Westerly	n	22966	140	11	1431	85	7
West Greenwich	n	5085	68	1	316	15	0
West Warwick	n	29581	426	34	2058	162	17
Woonsoket	y	43224	794	119	2872	213	22

Data in this page, Do Democrats prefer cats?

16.3: Data aggregation and correlation

Introduction

Correlations are easy to calculate, but interpretation beyond a strict statistical interpretation, e.g., two variables linearly associated, may be complicated — caution is recommended. With respect to interpreting a correlation, caution and temperance is warranted. As previously discussed, "**correlation is not causation**," is well known, but identifying when this applies to a particular analysis is not straight-forward. We introduced the problem of two variables sharing a **hidden covariation** which drives the correlation. In this section we introduce how correlations among grouped (aggregated) data may be quite different from the underlying individual correlations (cf. Robertson 1950, Greenland 2001, Portnov et. al 2006).

Data aggregation

Data aggregation or grouping refers to processes to group data in a summary form. Considerable public health data is presented this way. For example, the CDC reports table after table of data about morbidity and mortality of the United States of America population. Data are grouped by age, cities, counties, ethnicities, gender, and states and reports are generated to convey the status of health peoples. Similarly, education statistics, economic statistics, and statistics about crime are commonly crafted from grouped data of what originally was data for individuals.

Correlations between groups may yield spurious conclusions

Researchers interested in testing hypotheses like whether BMI is correlated with mortality (Flegal et al 2013, Kltasky et al 2017), or health disparities with ethnicity (Portnov et. al 2006), may use grouped data. In 16.2 we introduced the concept of **spurious correlation**. Correlations between grouped data may also mislead.

Consider the hypothesis that religiosity may deter criminal behavior. This hypothesis has been tested many times dating back to at least the 1940s (reviewed in Salvatore and Rubin 2018). Conclusions about religious beliefs range from negative association with criminal behavior to, in some reports, holding religious beliefs makes one more likely to commit crime. Testing versions of the hypothesis — what causes criminality in some individuals — among a variety of putative causal agents pops up through the history of biology research, arguably beginning with Galton. I hope you appreciate how challenging this would be to actually resolve — defining criminal behavior itself is laden with all kinds of sociology traps — and for a biologist, reeks of eugenics lore (Horgan 1993).

That all said, let's proceed to test the religion-criminality hypothesis with aggregated data. The null hypothesis would be no association between crime statistics and numbers of churches. We can also ask about association between crime and non-religious or secular beliefs. I added numbers of Catholic churches and secular humanists groups for cities larger than 100K population by Internet search (FBI for crime statistics, Wikipedia for cities). Figures 16.3.1 and 16.3.2 report crimes statistics aggregated by cities in the United States and by number of Catholic churches (Fig. 16.3.1), and by number of secular humanists groups (Fig. 16.3.2) in the same cities.

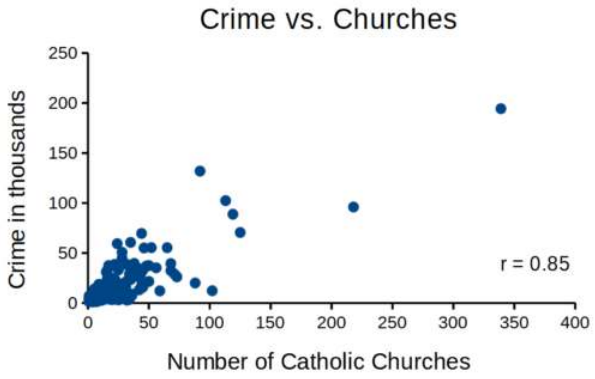


Figure 16.3.1: Scatterplot showing crime rates of cities by number of Catholic churches.

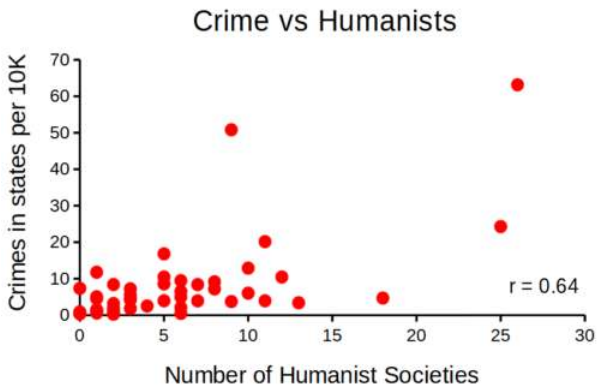


Figure 16.3.2: Scatterplot showing crime rates of cities by number of secular humanist associations.

We'll just take the numbers on faith (of course, we should think about the bunching around the origin — do we really think Internet search will get all of the secular groups, for example? Or is it really the case that several cities have no secular humanist groups?). Both correlations were statistically different from zero: crime by churches ($p < 0.001$) and crime by secular groups ($p < 0.001$).

Now, having read Chapter 16.2, I trust you recognize immediately that there's an important hidden covariate in common. Cities with small populations will have small numbers of crimes reported and smaller numbers of churches compared to large cities. Indeed, the correlation between population and crime for these cities was 0.89 and 0.97, respectively. However, after estimating the partial-correlations, we still have some explaining to do. For crime and churches, the partial correlation was $+0.37$ ($p = 0.009$); for crime and secular humanist groups, the partial correlation was -0.37 ($p = 0.018$). These results suggest that persons are more likely to commit crimes in cities with lots of Catholic churches whereas criminal behavior by individuals is less likely where secular humanist groups are numerous.

Before we start pointing fingers, the analysis presented here is a classic **ecological fallacy**. By grouping the data we lose information about the individuals, and it is the individuals to which the hypothesis applied. Thus, we are at risk of making incorrect conclusions by assuming that the individual is characterized by the group. The hypothesis remains challenging to test (how does one get a valid assessment of an individual's religiosity? The hypothesis is challenging to test, but studies of individuals tend to find no association or a negative association between criminal behavior and religiosity (Salvatore and Rubin 2018). Crime statistics may underestimate criminal behavior, e.g., embezzlement and other "white" crime), but a proper study would look to survey of individuals (Fig. 16.3.3).

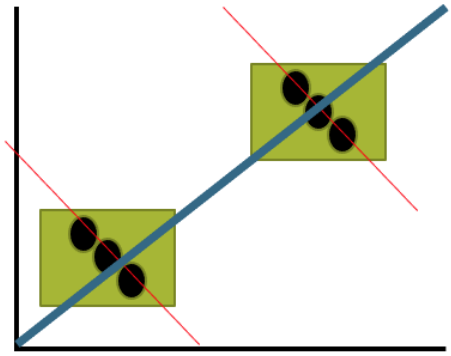


Figure 16.3.3: Illustration of ecological fallacy: positive association at level of groups (boxes, solid blue line), but negative association at level of individuals (black circles, red dashed lines).

Studies that use aggregate data test hypotheses about the groups, not about individuals in the groups. These studies are appropriate for comparing groups, e.g., health disparities by ethnicity (cite) or gender (cite), or comparisons among counties for medical resources (cite), but one cannot conclude that the association is present for members of the group.

Questions

[pending]

16.4: Spearman and other correlations

Introduction

Pearson product moment **correlation** is used to describe the level of linear association between two variables. There are many types of correlation estimators in addition to the familiar Product Moment Correlation, r.

Spearman rank correlation

If you take the ranks for X_1 and the ranks for X_2, the correlation of ranks is called **Spearman rank correlation**, r_s. Spearman correlation is a nonparametric statistic. Like the product moment correlation, it can take values between -1 and $+1$.

For variables X_1 and X_2, the rank order correlation may be calculated on the ranks as

$$\rho = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)}$$

where d_i is the difference between the ranks of X_1 and X_2 for each experimental unit. This formula assumes that there are no tied ranks; if there are, use the equation for the product moment correlation instead (but on the ranks).

R commander has an option to calculate the Spearman rank correlation simply by selecting the check box in the correlation sub menu. However, if the data set is small, it may be easier to just run the correlation in the script window.

Our example for the product moment correlation was between *Drosophila* fly wing length and wing area (Table 16.4.1).

Table 16.4.1. Fly wing lengths and area, units mm and mm^2, respectively (Dohm pers obs.)

Obs	Student	Length	Area
1	S01	1.524	0.446
2	S01	2.202	0.876
3	S01	1.52	0.39
4	S01	1.62	0.51
5	S01	1.71	0.736
6	S03	1.551	0.453
7	S03	2.228	0.882
8	S03	1.46	0.394
9	S03	1.659	0.503
10	S03	1.719	0.535
11	S05	1.534	0.441
12	S05	2.223	0.889
13	S05	1.49	0.391
14	S05	1.633	0.514
15	S05	1.704	0.546
16	S08	1.551	0.453
17	S08	2.228	0.882
18	S08	1.468	0.394
19	S08	1.659	0.503
20	S08	1.719	0.535

Data were collected by image analysis (ImageJ) of fixed wings to glass slides.

Here's the scatterplot of the ranks of fly wing length and fly wing area (Fig. 16.4.1).

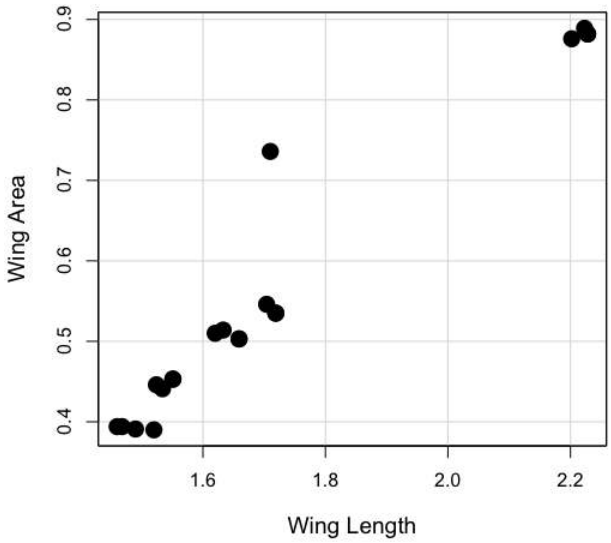


Figure 16.4.1: Drosophila wing area (mm^2) by wing length (mm).

A nonparametric alternative to the product moment correlation, the Spearman Rank correlation can be obtained directly. The Spearman correlation involves ranking the data, i.e., converting data types, from ratio scale data to ordinal scale, then applying the same formula used for the Product moment correlation to the ranked data. The Spearman correlation would be the choice for testing linear association between two ordinal type variables. It is also appropriate in lieu of the parametric product moment correlation when the statistical assumptions are not met, e.g., normality assumption.

R code

For the Spearman rank correlation, at the R prompt type

```
cor.test(Area, Length, alternative="two.sided", method="spearman")

R returns with
    Spearman's rank correlation rho

data:  Area and Length
S = 58.699, p-value = 5.139e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9558658
```

Alternatively, to calculate either correlation, use R Commander.

Rcmdr: Statistics → Summaries→ Correlation test

Example

```
BM=c(29,29,29,32,32,35,36,38,38,38,40)
Matings=c(0,2,4,4,2,6,3,3,5,8,6)

cor.test(BM,Matings, method="spearman")
```

```
Warning in cor.test.default(BM, Matings, method = "spearman") :
  Cannot compute exact p-value with ties

        Spearman's rank correlation rho

data:  BM and Matings
S = 77.7888, p-value = 0.03163
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6464143
```

```
cor.test(BM,Matings, method="pearson")

        Pearson's product-moment correlation

data:  BM and Matings
t = 2.6728, df = 9, p-value = 0.02551
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1087245 0.9042515
sample estimates:
      cor
0.6652136
```

Other correlations

Kendall's tau

Another nonparametric correlation is Kendall's tau (τ). Rank the X_1 values, then rank the X_2 values. Count the number of X_1, X_2 pairs that have the same rank (**concordant pairs**) and the number of X_1, X_2 pairs that do not have the same rank (**discordant pairs**), Kendall's tau is then

$$\tau = \frac{(no.\ of\ concordant\ pairs) - (no.\ of\ discordant\ pairs)}{\frac{1}{2}(n-1)}$$

where n is the number of pairs.

Note:

The denominator for τ is our familiar number of **pairwise comparisons** if we take $k = n$

We introduced concordant and discordant pairs when we presented McNemar's test and cross-classified experimental design in Chapter 9.6.

Example: Judging of Science Fair posters

What is the **agreement** between two judges, **A** and **B**, who evaluated the same science fair posters? Posters were evaluated on if the student's project was hypothesis-based and judges used a **Likert**-like scale Strongly disagree (1), Somewhat disagree (2), Neutral (3), Somewhat agree (4), Strongly agree (5).

Table 16.4.2. Two judges evaluated six posters for evidence of hypothesis-based project.

Poster	Judge.A	Judge.B

Poster	Judge.A	Judge.B
1	5	4
2	2	3
3	4	2
4	3	1
5	2	1
6	4	3

A concordant pair represents a poster ranked higher by both judges, while a **disconcordant pair** is a poster ranked high by one judge but low by another judge. Poster 1 and poster 5 were concordant pairs.

In R, it is simple to get this correlation directly by invoking the `cor.test` function and specifying the method equal to `kendall`. The `cor.test` assumes that the data are in a matrix, so use the `cbind` function to bind two vectors together – note the vectors need to have the same number of observations. If the data set is small, it is easier to just enter the data directly in the script window of R commander.

```
A = c(2,2,3,4,4,5)
B = c(1,3,1,2,3,4)
m = cbind(A,B)
cor.test(A,B, method="kendall")
  Cannot compute exact p-value with ties

        Kendall's rank correlation tau

data:  A and B
z = 1.4113, p-value = 0.1581
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5384615
```

End of R output

There were no ties in this data set, but we can run the product moment correlation just for comparison:

```
cor.test(A,B, method="pearson")

        Pearson's product-moment correlation

data:  A and B
t = 1.4649, df = 4, p-value = 0.2168
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4239715  0.9478976
sample estimates:
      cor
0.5909091
```

End R output

Tetrachoric and Polychoric correlation

Tetrachoric correlations used for binomial outcomes (yes, no), **polychoric correlation** used for ordinal categorical data like the Likert scale. Introduced by Karl Pearson, commonly applied correlation estimate for Item Analysis in psychometric research. Pyschometrics, a sub-discipline within psychology and now a significant part of education research, is about evaluating assessment tools.

R package `psych` .

R code: Tetrachoric correlation

```
tetrachoric(x,y=NULL,correct=.5,smooth=TRUE,global=TRUE,weight=NULL,na.rm=TRUE,
     delete=TRUE)
```

R code: Polychoric correlation

```
polychoric(x,smooth=TRUE,global=TRUE,polycor=FALSE,ML=FALSE, std.err=FALSE,
     weight=NULL,correct=.5,progress=TRUE,na.rm=TRUE,  delete=TRUE)
```

Polyserial correlation

R package `polychor` . Used to estimate linear association between a ratio scale variable and an ordinal variable.

R code: Polyserial correlation

```
polyserial(x,y)
```

Biserial correlation would be a special case of the polyserial correlation, where ordinal variable is replaced by a dichotomous (binomial) variable.

R code: Polyserial correlation

```
biserial(x,y)
```

Intra-class correlation coefficient

Both the ICC and the product moment correlation, r, which we introduced in Chapter 16.1, are measures of strength of linear association between two ratio scale variables (Jinyuan et al 2016). But ICC is more appropriate for association between **repeat measures** of the same thing, e.g., repeat measures of running speed. In contrast, the product moment correlation can be used to describe association between any two variables, e.g., between repeat measures of running speed, but also between say running speed and maximum jumping height. ICC is used when quantitative measures are organized into paired groups, e.g., before and after on same subjects, or cross-classified designs. ICC was introduced in Chapter 12.3 as part of discussion of repeated measures and random effects. ICC is used extensively to assess reliability of a measurement instrument (Shrout and Fleiss 1979; McGraw and Wong 1996).

Example. Data from Table 16.4.2

```
library(psych)
ICC(myJudge, lmer=FALSE)
```

R output follows

```
Intraclass correlation coefficients
                         type  ICC   F df1 df2      p lower bound upper bound
Single_raters_absolute   ICC1 0.40 2.3   5   6  0.166      -0.306        0.84
```

```
Single_random_raters     ICC2 0.46 3.9   5   5  0.081      -0.093        0.85
Single_fixed_raters      ICC3 0.59 3.9   5   5  0.081      -0.130        0.90
Average_raters_absolute ICC1k 0.57 2.3   5   6  0.166      -0.880        0.91
Average_random_raters   ICC2k 0.63 3.9   5   5  0.081      -0.205        0.92
Average_fixed_raters    ICC3k 0.74 3.9   5   5  0.081      -0.299        0.95

Number of subjects = 6 Number of Judges = 2
See the help file for a discussion of the other 4 McGraw and Wong estimates
```

Lots of output, lots of “ICC”. However, rather than explaining each entry, reflect on the type and review the data. Were the posters evaluated repeatedly? Posters were evaluated twice, but only once per judge, so there is a repeated design with respect to the posters. Were judges randomly selected from a population of all possible judges? No evidence was provided to suggest this, so judges were a fixed factor (see Chapter 12.3 and Chapter 14.3). The six ICC estimates reported by R follow discussion in Shrout and Fliess (1979), and our description fits their Case 3: “Each target is rated by each of the same k judges, who are the only judges of interest (p. 421)” Thus, we find ICC for single fixed rater, ICC = 0.59. Note that we would fail to reject the hypothesis that the judges evaluations were associated.

Questions

See Homework 9, Mike’s Workbook for biostatistics

16.5: Instrument reliability and validity

Introduction

What's in a **measure**? We've talked about measurement extensively, e.g., Chapter 3.3. As review, a measure is simply the result of some process used to quantify an object or item of interest. Instead of a number, a measure may return a classification: for a given sample unit, the unit may be categorized as meeting the definition and therefore given a "yes" or it does not ("no"), a dichotomous response. The method used to obtain the measurement is called the **instrument**. An instrument may indeed be an instrument like a sphygmomanometer or a thermocycler equipped with fluorescent optics. Instrument in this context, however, also includes questionnaires or surveys intended to determine people's responses on a particular topic.

In biology and biomedical research, there are thousands of kinds of measurements one has to choose among depending on the question at hand. In many cases choices are straightforward: in morphometrics, the instruments of choice will be lengths and areas and shapes quantified by rulers and application of well-defined geometry equations. Where multiple measurement approaches apply, **reliability analysis** can help decide which method to use, or, importantly, whether the different approaches agree. For example, Kruse et al (2017) compared ultra sound and magnetic resonance imaging measurements of Achilles tendon cross-sectional area; they found that although both methods were **internally consistent**, the methods consistently yielded different results.

In other arenas, the choice of instrument will be less clear. For example, doctors use a questionnaire to rank cardiac patients for attention in perioperative care, the care a surgical patient receives from admittance to release from the hospital, to improve patient outcomes. The questionnaire will include a number of questions intended to provide a summary picture of each patient so that if resources are limited, the most at risk patients may get priority. To the extent that the questionnaire in fact is a useful discriminant, then the instrument may benefit both hospital and patients.

In conducting measures one selects instruments that provide valid results. That is, provided the instruments are maintained and well-calibrated, use of a sphygmomanometer by a trained technician will return accurate and valid measures of a subject's blood pressure. Survey questions also can be evaluated for validity, although the extent to which survey questions measure what is intended may be more complex. For example, if the intent is to ascertain a subject's chance (i.e., risk) of graduating from college in the next year, how useful would the following question be if administered to a room filled with first-year students?

Survey question: How old are you now?

Simple enough question, but immediately, several questions come to mind. Do we want our responses in years, months, days, hours, minutes, or seconds? What about for those individuals that know only approximately when they were born (i.e., in many parts of the world, registration of birth is irregular)? So, we may even wonder about how necessary it is we ask this question of college students in the first place. E.g., do we really want to trigger our subjects for a case in which most of our subjects are about the same age?

Perhaps we decide this is important information to ask. When do you start counting? Most Western cultures start the clock at zero when the baby is born. In China and many other Eastern Asian countries, people are born at one. In India, once a person reaches a year plus six months, the person would be considered two years old, whereas in the USA, the person would still be considered one until the second birthday. Thus, depending on the person's culture identity, responses to this simple question may differ by as much as a year.

Types of reliability

Regardless of instrument, all measures contain error. Hence, even a valid instrument may not return an accurate measure for each subject. The concept of instrument **reliability** is concerned with error of measurement. Reliability may be defined in at least four contexts:

- internal consistency
- inter-rater (also called inter-observer reliability)
- parallel-forms
- test-retest

For an instrument to show internal consistency, this implies that the survey has multiple questions that pertain to the same concept or topic, but written in different ways to reveal effects of word choice, for example. Inter-rater reliability refers to an instrument that when used by different observers (e.g., science fair judges), the observers give the same or at least consistent scores for the

same test. Parallel forms reliability implies that two surveys on, perhaps, scientific literacy in high school students yield the same conclusions even though each survey has different questions. Test-retest reliability is a straight-forward concept — if the instrument is repeated, are the same scores achieved?

Reliability estimators

In general, correlation-type measures can be used to quantify the extent of reliability (also termed reproducibility or repeatability). The product moment correlation is used to quantify the relationship between two measures where there is clear distinction between the two variables. For example, to quantify the association between body weight and height, the proper correlation to calculate would be the product moment correlation because it is clear that a measure of weight goes with the variable weight whereas height measures goes with the variable height.

But it is less clear which variable should go first in the calculation when you have repeat measures of essentially the same thing. Which goes first, the first observation of sprint running speed over a 100 meters of the second measure of the same person's performance? Logically, we may say take the first as the X variable and the second as the Y variable, but there is no mathematical justification.

For a more challenging example, consider measures of body mass on male and female birds that are mated, and the researcher wants to assess whether there is a correlation between male and female weight — which variable goes first in the analysis, male weight or female weight? In such cases the intraclass correlation coefficient may be used (Chapter 12.3, Chapter 16.4). The intraclass correlation can be estimated as the ratio of the variance of interest over the sum of the variance of interest plus the variance error. Interest in the case of sprint running would be the two (or more) trials; for the bird weights, the variable of interest is weights of male and female birds within a mated pair. The formula for ICC in this context is given by

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_e^2} = \frac{MS_B - MS_e}{MS_B + (k-1)MS_e}$$

where k is the number of repeat measures, MS refers to the mean squares from the one-way ANOVA, B refers to variability between (among) subjects and e is the error or within-subjects variability.

Cronbach's alpha is a reliability measure that quantifies the internal consistency of items in a survey or instrument by calculating the average among these items. Cronbach's alpha will tend to increase as the intercorrelations among test items increase, and in this sense can be taken as an internal consistency estimate of the reliability of test scores.

$$Cronbach's\ \alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{n} s_i^2}{s_T^2}\right)$$

where k is the number of items, s_i^2 is the variance of the i-th item, and s_T^2 is the variance of the total score after summing all items.

Cronbach's alpha is one of the oldest measures, and at least in part because of how long ago it was introduced, is very common as a measure of reliability. However, there are other estimators and in some aspects these perform better than Cronbach's alpha.

Reliability statistics like Cronbach's alpha are available in the R package `psych`. See also R package `agRee`.

Example. Judging of posters

```
library(psych)
alpha(myJudge)
```

Output from R

```
Reliability analysis
Call: alpha(x = myJudge)

raw_alpha std.alpha G6(smc) average_r S/N   ase  mean   sd median_r
     0.74      0.74    0.59       0.59 2.9  0.21   2.8  1.1     0.59
```

```
lower alpha upper 95% confidence boundaries
0.33 0.74 1.15
```

Questions

[pending]

16.6: Similarity and distance

Introduction

A measure of dependence between two random variables. Unlike Pearson Product Moment correlation, **distance correlation** measures strength of association between the variables whether or not the relationship is linear. Distance correlation is a recent addition to the literature, first reported by Gábor J. Székely (e.g., Székely et al. 2007). The package `correlation` (Makowski et al 2019) offers distance correlation and significance test.

Example, fly wing dataset introduced 16.1 – Product moment correlation

```
library(correlation)
Area <- c(0.446, 0.876, 0.390, 0.510, 0.736, 0.453, 0.882, 0.394, 0.503, 0.535, 0.441, 0.889
Length <- c(1.524, 2.202, 1.520, 1.620, 1.710, 1.551, 2.228, 1.460, 1.659, 1.719, 1.534, 2.2
FlyWings <- data.frame(Area, Length)
correlation(FlyWings,method="distance")
```

Output from R

```
# Correlation Matrix (distance-method)
Parameter1 | Parameter2 |    r |       95% CI | t(169) |      p
------------------------------------------------------------------
      Area |     Length | 0.92 | [0.80, 0.97] |  30.47 | < .001***

p-value adjustment method: Holm (1979)
Observations: 20
```

The product-moment correlation was 0.97 with 95% confidence interval (0.92, 0.99). The note about "p-value adjustment method: Holm (1979)" refers to the algorithm used to mitigate the **multicomparison problem**, which we first introduced in Chapter 12.1. The correction is necessary in this context because of how the algorithm conducts the test of the distance correlation. Please see Székely and Rizzo (2013) for more details.

Which should you report? For cases where it makes sense to test for a linear association, then the product-moment correlation is the one to use. For other cases where no inference of linearity is expected, then the distance correlation makes sense.

Similarity and Distance

Similarity and distance are related mathematically. When two things are similar, the distance between them is small; When two things are dissimilar, the distance between them is great. Whether similarity (sometimes **dissimilarity**) or distance, the estimate is a statistic. The difference between the two is that the typical distance measures one sees in biostatistics all obey the **triangle inequality rule** while similarity (dissimilarity) indices do not necessarily obey the triangle inequality rule.

Distance measures

Distance is a way to talk about how far (or how close) two objects are from each other (Fig. 16.6.1). The distance may be relate to physical distance (**map distance**), or in mathematics, distance is a metric or statistic. **Euclidean distance** is the distance between two points in either the xy-plane or 3-dimensional space measures the length of a segment connecting the two points (e.g., $x_1, y_1 = 1, 4$ and $x_2, y_2 = 1, 4$).

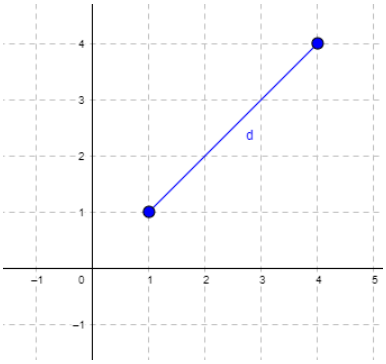


Figure 16.6.1: **Cartesian plot** of two points, the first at $x_1 = 1$ and $y_1 = 1$ and the second at $x_2 = 4$ and $y_2 = 4$.

For two points $(x_1, y_1$ and $x_2, y_2)$ described in two dimensions (e.g., an xy-plane), the distance d is given by

$$d = \sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2}$$

For two points described in three (e.g., an xyz-space), or more dimensions, the distance d is given by

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Distances of this form are Euclidean distances and can be directly obtained by use of the **Pythagorean Theorem**. The **triangle inequality rule** then applies (i.e., the sum of any two sides must be less than the length of the remaining side). Euclidean distance measures also include

- **Manhattan distance:** the sum of absolute difference between the measures in all dimensions of two points.

  $|x_1 - x_2| + |y_1 - y_2|$
- **Chebyshev distance:** also called the maximum value distance, the distance between two points is the greatest of their differences along any coordinate dimension.

  $max\left(|x_1 - x_2|, |y_1 - y_2|\right)$

Note: We first met Chebyshev in Chapter 3.5.

Example

There are a number of distance measures. Let's begin discussion of distance with geographic distance as an example. Consider the distances between cities (Table 16.6.1).

Table 16.6.1. Distances (miles) among cities.

	Honolulu	Seattle	Manila	Tokyo	Houston
Honolulu	0	2667.57	5323.37	3849.99	3891.82
Seattle	2667.57	0	6590.23	4776.81	1888.06
Manilla	5323.37	6590.23	0	1835.1	8471.48
Tokyo	3849.99	4776.81	1835.1	0	6664.82
Houston	3891.82	1888.06	8471.48	6664.82	0

This table is a **distance matrix** — note that along the diagonal are "zeros," which should make sense — the distance between an object and itself is, well, zero. Above and below the diagonal you see the distance between one city and another. This is a special kind of matrix called a **symmetric matrix**. Enter the distance in miles (1 mile = 1.609344) between 2 cities (this is "**pairwise**"). There are many resources "out there," to help you with this. For example, I found a web site called mapcrow that allowed me to enter the cities and calculate distances between them.

To get the distance matrix, use this online resource, the Geographic Distance Matrix Calculator.

For a real-world problem, use `geodist` package. Provide latitude and longitude coordinates.

Distance measures used in biology

It is easy to see how the concept of **genetic distance** between a group of species (or populations within a species) could be used to help build a network, with genetically similar species grouped together and genetically distant species represented in a way to represent how far removed they are from each other. Here, we speak of distance as in similarity: two species (populations) that are similar are close together, and the distance between them is short. In contrast, two species (populations) that are not similar would be represented by a great distance between them. Genetic distance is the amount of divergence of species from each other. Smaller genetic distances reflects close genetic relationship.

Here's an example (Fig. 16.6.2), RAPD gel for five kinds of beans. RAPD stands for random amplified polymorphic DNA.

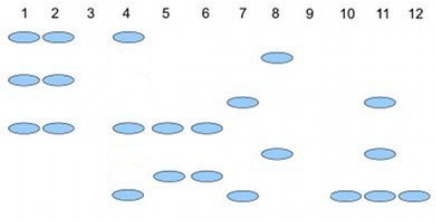


Figure 16.6.2: RAPD gel (simulated) five kinds of beans.

Samples were small red beans (SRB), garbanzo beans (GB), split green pea (SGP), baby lima beans (BLB), and black eye peas (BEP). RAPD primer 1 was applied to samples in lanes 1 – 6; RAPD primer 2 was applied to samples in lane 7 – 12. Lane 1 & 7 = SRB; Lane 2 & 8 = GB; Lanes 3 & 9 = SGP; Lane 4 & 10 = BLB; Lane 5 & 11 = BB; Lane 6 & 12 = BEP.

Here's how to go from gel documentation to the information needed for genetic distance calculations (see below). I'll use "1" to indicate presence of a band, "0" to indicate absence of a band, and "?" to indicate no information. For simplicity, I ignored the RF value, but ranked the bands by order of largest (= 1) to smallest (=8) fragment.

We need three pieces of information from the gel to calculate genetic distance.

N_A = the number of markers for taxon A

N_B = the number of markers for taxon B

N_{AB} = the number of markers in common between A and B (this is the pairwise part — we are comparing taxa two at a time).

First, compare the beans against the same primer. My results for primer 1 are in Table 16.6.2; results for primer 2 are in Table 16.6.3.

Table 16.6.2. Bands for Primer 1

marker	lane 1	Lane 2	Lane3	Lane 4	Lane 5	Lane 6
1	1	1	?	1	0	0
3	1	1	?	0	0	0
5	1	1	?	1	1	1
7	0	0	?	0	1	1

Table 16.6.3. Bands for Primer 2

marker	Lane 7	Lane 8	Lane 9	Lane 10	Lane 11	Lane 12
2	0	1	?	0	0	0
4	1	0	?	0	1	0
6	0	1	?	0	1	0
8	1	0	?	1	1	1

From Table 16.6.2 and Table 16.6.3, count \(N_{A} (= N_{B})\) for each taxon. Results are in Table 16.6.4

Table 16.6.4. Bands for each taxon.

Taxon	No. markers from Primer1	No. markers from Primer2	Total
SRB	3	2	5

GB	3	2	5
SGP	?	?	?
BLB	2	1	3
BB	2	3	5
BEP	2	1	3

As you can see, there is no simple relationship among the taxa; there is no obvious combination of markers that readily group the taxa by similarity. So, I need a computer to help me. I need a measure of genetic distance, a measure that indicates how (dis)similar the different varieties are for our genetic markers. I'll use a distance calculation that counts only the "present" markers, not the "absent" markers, which is more appropriate for RAPD. I need to get the N_{AB} values, the number of shared markers between pairs of taxa.

Table 16.6.5. N_{AB} values.

	SRB	GB	BLB	BB	BEP
SRB	0	3	3	3	2
GB		0	2	2	1
BLB			0	2	2
BB				0	3
BEP					0

The equation for calculating **Nei's distance** is:

$$Nei's\, d = 1 - \left(\frac{N_{AB}}{N_A + N_B - N_{AB}} \right)$$

where N_A = number of bands in taxon "A", N_B = number of bands in taxon "B", and N_{AB} is the number of bands in common between A and B (Nei and Li 1979). Here's an example calculation.

Let A = SRB and B = GB, then

$$Distance = 1 - \left(\frac{3}{5 + 5 - 3} \right) = 0.5714$$

Questions

1. Review all of the different correlation estimates we have introduced in Chapter 16 and construct a table to help you learn. Product moment correlation is presented as example.

Name of correlation	variable 1 type	variable 2 type	purpose
Product moment	ratio	ratio	estimate linear association

16.7: References and suggested readings

Bartko, J. (1976). On various intraclass reliability coefficients. *Psychological Bulletin* 83:762-765.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57:289-300.

Brand, J., Altman, D. (1997). Statistics notes: Cronbach's alpha. *BMJ* 314:572. https://doi.org/10.1136/bmj.314.7080.572.

Bruton, A., Conway, J. H., and Holgate, S. T. (2000). Reliability: What is it and how is it measured? *Physiotherapy* 86:94-99.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297-334.

Doornick, J. A., Hansen H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* 70(S1):927–939

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 527-537.

Greenland, S. (2001). Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International journal of epidemiology*, *30*(6), 1343-1350.

Hu, P., Jiao, R., Jin, L., & Xiong, M. (2018). Application of Causal Inference to Genomic Analysis: Advances in Methodology. *Frontiers in Genetics*, *9*, 238. https://doi.org/10.3389/fgene.2018.00238.

Kleinberg, S., & Hripcsak, G. (2011). A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, *44*(6), 1102–1112. https://doi.org/10.1016/j.jbi.2011.07.001.

Kruse, A., Stafilidis, S., & Tilp, M. (2017). Ultrasound and magnetic resonance imaging are not interchangeable to assess the Achilles tendon cross-sectional-area. *European Journal of Applied Physiology*, *117*(1), 73-82.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59-66.

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2019). Methods and Algorithms for Correlation Analysis in R. Journal of Open Source Software, 5(51), 2306. https://doi.org/10.21105/joss.02306

Nei, M., Li ,W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS* 76:5269-5273. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC413122/

Parascandola, M., & Weed, D. L. (2001). Causation in epidemiology. Journal of Epidemiology & Community Health, 55(12), 905–912.

Portnov, B. A., Dubnov, J., & Barchana, M. (2007). On ecological fallacy, assessment errors stemming from misguided variable selection, and the effect of aggregation on the outcome of epidemiological study. *Journal of exposure science & environmental epidemiology*, *17*(1), 106-121.

Robinson, W.S. Ecological correlations and the behavior of individuals. *Am Sociol Rev* 1950:15:351–357.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment* 8:350-353.

Shrout, P., Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86:420-428.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74:107-120.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6), 2769-2794.

U.S. Pet Ownership & Demographics Sourcebook (2012) American Veterinary Association. Link to pdf file

Wetzels, R. & Wagenmakers, E.-J. (2012) A default Bayesian hypothesis test for correlations and partial correlations. *Psychon Bull Rev* 19: 1057-1064

Wilson, C. (2014). It's True: Liberals Like Cats More Than Conservatives Do. Time Magazine online (https://time.com/8293/its-true-liberals-like-cats-more-than-conservatives-do/)

Zou, K. H., Tuncali, K., Silverman, S. G. (2003). Correlation and linear regression. *Radiology* 227(3): 617-628.

CHAPTER OVERVIEW

17: Linear Regression

Introduction

Regression is a toolkit for developing **models** of **cause and effect** between one **ratio scale** data type **dependent** response variables, and one (**simple linear regression**) or more or more (**multiple linear regression**) ratio scale data type **independent** predictor variables. By convention the dependent variable(s) is denoted by Y, the independent variable(s) represented by $X_1, X_2, \ldots X_n$ for n independent variables. Like ANOVA, linear regression is simply a special case of the **general linear model**, first introduced in Chapter 12.7.

Components of a statistical model

Regression statistical methods return **model** estimates of the intercept and slope **coefficients**, plus statistics of regression fit (e.g., R^2, aka "R-squared," the **coefficient of determination**).

Chapter 17.1 – 17.9 cover the simple linear model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Chapter 18.1 – 18.5 cover the multiple regression linear model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon_i$$

where α or β_0 represent the **Y-intercept** and β or $\beta_1, \beta_2, \ldots \beta_n$ represent the regression **slopes**.

Regression and correlation test linear hypotheses

We state that the relationship between two variables is linear (the **alternate hypothesis**) or it is not (the null hypothesis). The difference? **Correlation** is a test of linear association (are variables correlated, we ask?), imply possible **causation**, but are not sufficient evidence for causation: we do not imply that one variable causes another to vary, even if the correlation between the two variables is large and positive, for example. Correlations are used in statistics on data sets not collected from explicit experimental designs incorporated to test specific hypotheses of cause and effect.

Linear regression, however, is to cause and effect as correlation is to association. With regression and ANOVA, we are indeed making a case for a particular understanding of the cause of variation in a response variable: modeling cause and effect is the goal. Regression, ANOVA, and other general linear models are designed to permit the statistician to control for the effects of **confounding variables** provided the causal variables themselves are uncorrelated.

Assumptions of linear regression

The key assumption in linear regression is that a straight line indeed is the best fit of the relationship between dependent and independent variables. The additional assumptions of parametric tests (Chapter 13) also hold. In Chapter 18 we conclude with an extension of regression from one to many predictor variables and the special and important topic of correlated predictor variables or **multicollinearity**.

Build a statistical model, make predictions

In our exploration of linear regression we begin with simple linear regression, also called ordinary least squares regression, starting with one predictor variable. Practical aspects of model diagnostics are presented. Regression may be used to describe or to provide a predictive statistical framework. In Chapter 18 we conclude with an extension of regression from one to many predictor variables. We conclude with a discussion of model selection. Throughout, use of `Rcmdr` and R have multiple ways to analyze linear regression models are presented; we will continue to emphasize the general linear model approach, but note that use of linear model in `Rcmdr` provides a number of default features that are conveniently available.

References

Linear regression is a huge topic; references I include are among my favorite on the subject, but are only a small and incomplete sampling. For simplicity, I merged references for Chapter 17 and Chapter 18 into one page at References and suggested readings

(Ch17 & 18)

17.1: Simple linear regression

Introduction

Linear regression is a toolkit for developing **linear models** of **cause and effect** between a ratio scale data type, response or **dependent variable**, often labeled Y, and one or more ratio scale data type, predictor or independent variables, X. Like ANOVA, linear regression is a special case of the **general linear model**. Regression and correlation both test linear hypotheses: we state that the relationship between two variables is linear (the **alternate hypothesis**) or it is not (the **null hypothesis**). The difference?

- **Correlation** is a test of association (are variables correlated, we ask?), but are not tests of causation: we do not imply that one variable causes another to vary, even if the correlation between the two variables is large and positive, for example. Correlations are used in statistics on data sets not collected from explicit experimental designs incorporated to test specific hypotheses of cause and effect.
- Linear regression is to cause and effect as correlation is to association. With regression and ANOVA, which again, are special cases of the general linear model (LM), we are indeed making a case for a particular understanding of the cause of variation in a response variable: modeling cause and effect is the goal.

We start our LM model as $Y \sim model$ where "$\sim$", **tilda**, is an operator used by R in formulas to define the relationship between the response variable and the predictor variable(s).

From R Commander we call the linear model function by **Statistics → Fit models → Linear model …** , which brings up a menu with several options (Fig. 1).

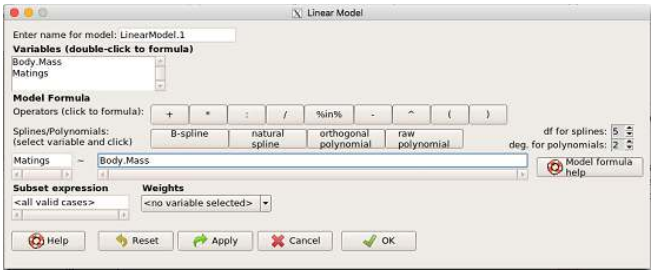


Figure 17.1.1: R commander menu interface for linear model.

Our model was

```
Matings ~ Body.Mass
```

R commander will keep track of the models created and enter a name for the object. You can, and probably should, change the object name yourself. The example shown in Figure 17.1.1 is a simple linear regression, with `Body.Mass` as the Y variable and `Matings` the X variable. No other information need be entered and one would simply click OK to begin the analysis.

Example

The purpose of regression is similar to ANOVA. We want a model, a statistical representation to explain the data sample. The model is used to show what causes variation in a response (dependent) variable using one or more predictors (independent variables). In life history theory, mating success is an important trait or characteristic that varies among individuals in a population. For example we may be interested in determining the effect of age (X_1) and body size (X_2) on mating success for a bird species. We could handle the analysis with ANOVA, but we would lose some information. In a clinical trial, we may predict that increasing Age (X_1) and BMI (X_1) causes increase blood pressure (Y).

Our causal model looks like $X_1 + X_2 \rightarrow Y$.

Let's review the case for ANOVA first.

The response (dependent variable), the number of successful matings for each individual, would be a quantitative (interval scale) variable. (Reminder: You should be able to tell me what kind of analysis you would be doing if the dependent variable was categorical!) If we use ANOVA, then factors have levels. For example, we could have several adult birds differing in age (factor 1) and of different body sizes. Age and body size are quantitative traits, so, in order to use our standard ANOVA, we would have to assign individuals to a few levels. We could group individuals by age (e.g., < 6 months, 6 – 10 months, > 12 months) and for body

size (e.g., small, medium, large). For the second example, we might group the subjects into age classes (20-30, 30-40, etc), and by AMA recommended BMI levels (underweight < 18.5, normal weight 18.5 – 24.9, overweight 25-29.9, obese > 30).

We have not done anything wrong by doing so, but if you are a bit uneasy by this, then your intuition will be rewarded later when we point out that in most cases you are best to leave it as a quantitative trait. We proceed with the test of ANOVA, but we are aware that we've lost some information — continuous variables (age, body size, BMI) were converted to categories — and so we suspect (correctly) that we've lost some power to reject the null hypothesis. By the way, when you have a "factor" that is a continuous variable, we call it a "**covariate**." Factor typically refers to a categorical explanatory (independent) variable.

We might be tempted to use correlation — at least to test if there's a relationship between Body Mass and Number of Matings. Correlation analysis is used to measure the intensity of association between a pair of variables. Correlation is also used to to test whether the association is greater than that expected by chance alone. We do not express one as causing variation in the other variable, but instead, we ask if the two variables are related (**covary**). We've already talked about some properties of correlation: it ranges from -1 to $+1$ and the null hypothesis is that the true association between two variables is equal to zero. We will formalize the correlation next time to complete our discussion of the linear relationship between two variables.

But regression is appropriate here because we are indeed making a causal claim: we selected Age and Body Size, and we selected Age and BMI in the second example wish to develop a model so we can predict and maybe even advise.

Least squares regression explained

Regression is part of the general linear model family of tests. If there is one linear predictor variable, then that is a **simple linear regression (SLR)**, also called **ordinary least squares (OLS)**, if there are two or more linear predictor variables, then that is a **multiple linear regression** (MLR, Chapter 18).

First, consider one predictor variable. We begin by looking at how we might summarize the data by fitting a line to the data; we see that there's a relationship between mass and mating success in both young and old females (and maybe in older males).

The data set was

Table 17.1.1. Our data set of number of matings by male bird by body mass (g).

ID	Body.Mass	Matings
1	29	0
2	29	2
3	29	4
4	32	4
5	32	2
6	35	6
7	36	3
8	38	3
9	38	5
10	38	8
11	40	6

And a **scatterplot** of the data (Fig. 17.1.2)

Figure 17.1.2: Number of matings by body mass (g) of the male bird.

There's some scatter, but our eyes tell us that as body size increases, the number of matings also increases. We can go so far as to say that we can predict (imperfectly) that larger birds will have more matings. We fit the **best-fit line** to the data and added the line to our scatterplot (Fig. 17.1.3). The best-fit line meets the requirements that the error about the line is minimized (see below). Thus, we would predict about six matings for a 40-gram bird, but only two matings for a 28-gram bird. And this is a good feature of regression, prediction, as long as used with some caution.

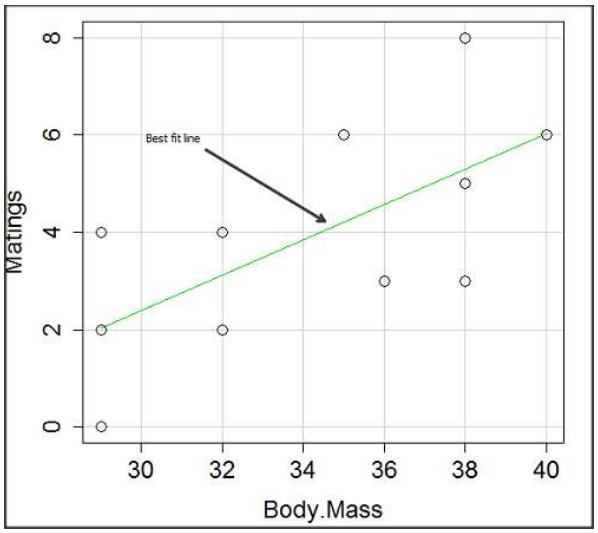


Figure 17.1.3: Same data as in Fig. 17.1.2, but with the “best fit” line.

Note that prediction works best for the range of data for which the regression model was built. Outside the range of values, we predict with caution.

The simplest linear relationship between two variables is the SLR. This would be the parameter version (population, not samples), where $Y_i = \alpha + \beta X_i + \epsilon_i$

α = the **Y-intercept coefficient** and it is defined as $a = \bar{Y} - b\bar{X}$. Solve for intercept by setting $X = 0$.

β = the **regression coefficient** (slope)

β = \dfrac{\sum \left(X_{i} - \bar{X}\right) \cdot \left(Y_{i} - \bar{Y}\right)}{\sum \left(X_{i} - \bar{X}\right)^{2}}\)

Note that the denominator is just our corrected sums of squares that we’ve seen many times before. The numerator is the cross-product and is referred to as the covariance.

ϵ = the error or “residual”, $\epsilon_i = Y_i - \hat{Y}_i$

The **residual** is an important concept in regression. We briefly discussed “what’s left over,” in ANOVA, where an observation Y_i is equal to the population mean plus the factor effect of level i plus the remainder or “error”.

In regression, we speak of residuals as the departure (difference) of an actual Y_i (observation) from the predicted Y ($\hat{Y}$, say “Y-hat”).

The linear regression predicts Y, and what remains unexplained by the regression equation is called the residual. There will be as many residuals as there were observations.

But why THIS particular line? We could draw lines anywhere through the points. Well, this line is termed the **“**best fit**“** because it is the only line that minimizes the sum of the squared deviations for all values of Y (the observations) and the predicted $\hat{Y}$. The best fit line minimizes the sum of the squared residuals, $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$.

Thus, like ANOVA, we can account for the total sums of squares (SS_{tot}) as equal to the **sums of squares** (variation), explained by the regression model, (SS_{reg}), plus what’s not explained, what’s left over, the residual sums of squares, (SS_{res}), aka (SS_{error}).

$$SS_{tot} = SS_{reg} + SS_{res}$$

Models used to predict new values

Once a line has been estimated, one use is to **predict** new observations not previously measured!

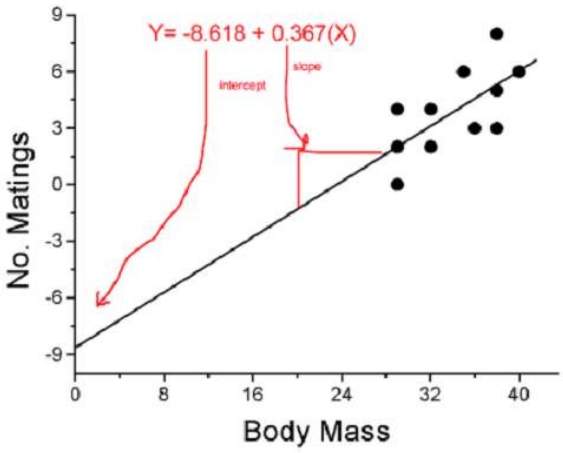


Figure 17.1.4: Figure 17.1.3 redrawn to extend the line to the Y-intercept.

This is an important use of models in statistics: use an equation to fit to some data, then predict Y values from new values of X. To use the equation, simply insert new values of X into the equation, because the slope and intercept are already "known." Then for any X_i we can determine $\hat{Y}$ (predicted Y value that is on the best-fit regression line).

This is what people do when they say

> "if you are a certain weight (or BMI) you have this increased risk of heart disease"
>
> "if you have this number of black rats in the forest you will have this many nestlings survive to leave the nest"
>
> "if you have this much run-off pollution into the ocean you have this many corals dying"
>
> "if you add this much enzyme to the solution you will have this much resulting product".

R Code

We can use the drop down menu in `Rcmdr` to do the bulk of the work, supplemented with a little R code entered and run from the script window. Scrape data from Table 17.1.1 and save to R as `bird.matings` .

```
LinearModel.3 <- lm(Matings ~ Body.Mass, data=bird.matings)
summary(LinearModel.3)
```

Output from R:

```
Call:
lm(formula = Matings ~ Body.Mass, data = bird.matings)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29237 -1.34322 -0.03178 1.33792 2.70763

Coefficients:
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  -8.4746     4.6641   -1.817    0.1026
Body.Mass     0.3623     0.1355    2.673    0.0255 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.776 on 9 degrees of freedom
```

```
Multiple R-squared: 0.4425, Adjusted R-squared: 0.3806
F-statistic: 7.144 on 1 and 9 DF, p-value: 0.02551
```

Get the sum of squares from the **ANOVA table**

```
myAOV.full <- anova(LinearModel.3); myAOV.full
```

Output from R, the ANOVA table

```
Analysis of Variance Table

Response: Matings
           Df  Sum Sq  Mean Sq  F value     Pr(>F)
Body.Mass   1  22.528  22.5277   7.1438   0.02551 *
Residuals   9  28.381   3.1535
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can do more,

```
str(myAOV.full)
```

Note:

`str()` command lets us look at an object created in R. Type `?str` or `help(str)` to bring up the R documentation. Here, we use `str()` to look at the structure of the object we created, `myAOV.full` . "Classes" refers to the R programming `class` attribute inherited by the object.

Output from R:

```
Classes 'anova' and 'data.frame': 2 obs. of 5 variables:
$ Df : int 1 9
$ Sum Sq : num 22.5 28.4
$ Mean Sq: num 22.53 3.15
$ F value: num 7.14 NA
$ Pr(>F) : num 0.0255 NA
- attr(*, "heading")= chr [1:2] "Analysis of Variance Table\n" "Response: Matings"
```

Extract the sum of squares: type the object name then `$"Sum Sq"` at the R prompt.

```
myAOV.full $"Sum Sq"
```

Output from R:

```
[1] 22.52773 28.38136
```

Get the residual sum of squares.

```
SSE.myAOV.full <- myAOV.full $"Sum Sq"[2]; SSE.myAOV.full
```

Output from R:

```
[1] 22.52773
```

Get the regression sum of squares.

```
SSR.myAOV.full <- myAOV.full $"Sum Sq"[1]; SSR.myAOV.full
```

Output from R:

```
[1] 50.90909
```

Now, get the total sums of squares for the model.

```
ssTotal.myAOV.full <- SSE.myAOV.full + SSR.myAOV.full; ssTotal.myAOV.full
```

Calculate the coefficient of determination.

```
myR_2 <- 1 - (SSE.myAOV.full/(ssTotal.myAOV.full)); myR_2
```

Output from R:

```
[1] 0.4425091
```

Which matches what we got before, as it should.

Regression equations may be useful to predict new observations

True. However, you should avoid making estimates beyond the range of the X-values that were used to calculate the best-fit regression equation! Why? The answer has to do with the shape of the confidence interval around the regression line.

I've drawn an exaggerated **confidence interval (CI),** for a regression line between an X and a Y variable. Note that the CI is narrow in the middle, but wider at the end. Thus, we have more confidence in predicting new Y values for data that fall within the original data because this is the region where we are most confident.

Calculating the CI for the linear model follows from CI calculations for other estimates. It is a simple concept — both the intercept and slope were estimated with error, so we combine these into a way to generalize our confidence in the regression model as a whole given the error in slope and intercept estimation.

$$95\%\ CI = b_1 \pm t_{df} SE_{b1}$$

The calculation of confidence interval for the linear regression involves the standard error of the residuals, the sample size, and expressions relating the standard deviation of the predictor variable X — we use the t-distribution.

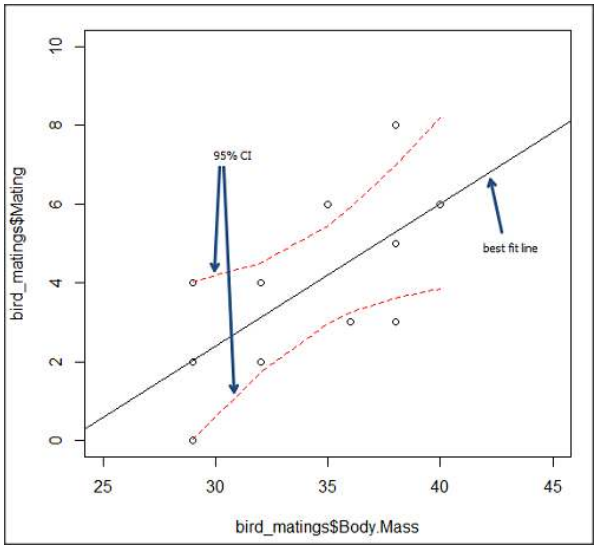


Figure 17.1.5: 95% confidence interval about the best fit line.

How I got this graph

```
plot(bird_matings$Body.Mass,bird_matings$Mating,xlim=c(25,45),ylim=c(0,10))
mylm <- lm(bird_matings$Mating~bird_matings$Body.Mass)
predict(mylm, interval = c("confidence"))
abline(mylm, col = "black")
x<-bird_matings$Body.Mass
lines(x, prd[,2], col= "red", lty=2)
lines(x, prd[,3], col= "red", lty=2)
```

Nothing wrong with my code, but getting all of this to work in R might best be accomplished by adding another package, a plug-in for Rcmdr called `RcmdrPlugin.HH` . HH refers to Heiberger and Holland, who designed this package specializing in graphical displays of data and data analysis.

Assumptions of OLS, introduction

We will cover assumptions of OLS in detail in 17.8, Assumptions and model diagnostics for Simple Linear Regression. For now, briefly, the **assumptions** for OLS regression include:

1. **Linear** model is appropriate: the data are well described (fit) by a linear model
2. **Independent** values of Y and equal variances. Although there can be more than one Y for any value of X, the Y's cannot be related to each other (that's what we mean by independent). Since we allow for multiple Y's for each X, then we assume that the variances of the range of Y's are equal for each X value (this is similar to our ANOVA assumptions for equal variance by groups).
3. **Normality**. For each X value there is a normal distribution of Y's (think of doing the experiment over and over)
4. **Error** (residuals) are normally distributed with a mean of zero.

Additionally, we assume that measurement of X is done without error (the equivalent, but less restrictive practical application of this assumption is that the error in X is at least negligible compared to the measurements in the dependent variable). Multiple regression makes one more assumption, about the relationship between the predictor variables (the X variables). The assumption is that there is no **multicolinearity**: the X variables are not related or associated to each other.

In some sense the first assumption is obvious if not trivial — of course a "line" needs to fit the data so why not plow ahead with the OLS regression method, which has desirable statistical properties and let the estimation of slopes, intercept and fit statistics guide us? One of the really good things about statistics is that you can readily test your intuition about a particular method using data simulated to meet, or not to meet, assumptions.

Coming up with datasets like these can be tricky for beginners. Thankfully others have stepped in and provide tools useful for data simulations which greatly facilitate the kinds of testing of assumptions statisticians greatly encourage us all to do (see Chatterjee and Firat 2007).

Questions

1. True or False. Regression analysis results in a model of the cause-effect relationship between a dependent and one (simple linear) or more (multiple) predictor variables. The equation can be used to predict new observations of the dependent variable.
2. True or False. The value of X at the Y-intercept is always equal to zero in a simple linear regression.
3. **Anscombe's quartet** (Anscombe 1973) is a famous example of this approach and the fictitious data can be used to illustrate the fallacy of relying solely on fit statistics and coefficient estimates.

   Here are the data (modified from Anscombe 1973, p. 19) — I leave it to you to discover the message by using linear regression on Anscombe's data set. Hint: play naïve and generate the appropriate descriptive statistics, make scatterplots for each X, Y set, then run regression statistics, first on each of the X, Y pairs (there were four sets of X, Y pairs).

Set 1		**Set 2**		**Set 3**		**Set 4**	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

17.2: Relationship between the slope and the correlation

Introduction

Product moment correlation is used to indicate the strength of the linear association between two ratio-scale variables; the **slope** tells you the rate of change between the two variables. When the correlation is negative, the slope will be negative; when correlation is positive, so too will the slope.

As you might suspect, there is a mathematical relationship between the product moment correlation, r, and the regression slope, b_1. We haven't spent much time explaining the equations presented in this text, but correlation and linear regression are such important tools it's worth a closer look.

Recall the equation of the correlation is

$$r_{XY} = \frac{\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{(n-1)s_X s_Y}$$

where the numerator is termed the **covariance** between ***X*** and ***Y*** and the denominator contains the standard deviations of ***X*** and ***Y*** variables. We can say the at the covariance is standardized by the variability in ***X*** and ***Y***. In contrast, the regression slope is equal to the covariance divided by the variance in ***X***.

$$b_1 = \frac{\sum_{i=1}^{n}\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sum_{i=1}^{n}(n-1)s_X s_Y}$$

Thus, with a little algebra, we can see that the slope and correlation are equal to each other as

$$b_1 = r \cdot \frac{s_X}{s_Y}$$

This should drive home the following **statistical reasoning** point. You can always calculate a slope from a correlation, but recall that correlation analysis is intended as a test of the hypothesis of a **linear association** between variables for which **cause and effect** model — though perhaps reasonable — should not always be implied. Just because it is mathematically possible does not mean the analysis is correct for the problem.

Questions

1. If the correlation is 0.6, $s_{\bar{X}} = 2.3$, and $s_{\bar{Y}} = 1.67$, what is the slope?

17.3: Estimation of linear regression coefficient

Introduction

In discussing correlations we made the distinction between inference, testing the statistical significance of the estimate, and the process of getting the estimate of the parameter itself. **Estimating parameters** is possible for any data set; whether or not the particular model is a good and useful model is another matter. Statisticians speak about the **fit of a model**… that a model with one or more **independent predictor variables** explains a substantial amount of the variation in the **dependent variable**, that it describes the relationship between the predictors and the dependent variable without bias. A number of tools have been developed to assess model fit. For starters, I'll list just two ways you can approach whether a **linear model** fits your data or requires some intervention on your part.

Assess fit of a linear model

Recall our R output from the regression of Number of Matings on Body Mass from the bird data set. We used the linear model function.

```
LinearModel.1 <- lm(Matings ~ Body.Mass, data=bird_matings)
summary(LinearModel.1)
Call: lm(formula = Matings ~ Body.Mass, data = bird_matings)
Residuals:       Min       1Q    Median       3Q      Max
            -2.29237 -1.34322 .-0.03178..1.33792 .2.70763
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -8.4746      4.6641   -1.817    0.1026
Body.Mass      0.3623      0.1355    2.673    0.0255 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.776 on 9 degrees of freedom
Multiple R-squared: 0.4425, Adjusted R-squared: 0.3806
F-statistic: 7.144 on 1 and 9 DF, p-value: 0.02551
```

Request R to print the ANOVA table.

```
Anova(RegModel.1, type="II")
Anova Table (Type II tests)
Response: Matings
            Sum Sq   Df   F value    Pr(>F)
Body.Mass   22.528    1    7.1438   0.02551 *
Residuals   28.381    9
```

With a little rounding we have the following statistical model:

$$Y_i = -8.5 + 0.36 \cdot X_i$$

and in English, *Number of matings equals* `Body.mass` *multiplied by 0.36 then subtract 8.5;* the intercept was -8.5, the slope was 0.36.

Note:

The results of the regression analyses have been stored in the object called " `LinearModel.1` ". This is a nice feature of Rcmdr — it automatically provides an object name for you. Note that with each successive run of the linear model function via Rcmdr that it will change the object name by adding numbers successively. For example, after `LinearModel.1` the next

run of `lm()` in Rcmdr will automatically be called " `LinearModel.2` " and so on. In your own work you may specify the names of the objects directly or allow Rcmdr to do it for you, but do keep track of the object names!

From the R output we see that the estimate of the slope was +0.36, statistically different from zero ($p = 0.025$). The intercept was -8.5, but not statistically significant ($p = 0.103$), which means the intercept may be zero.

As a general rule, if you make an estimate of a parameter or coefficient, then you should provide a confidence interval in the following form:

$$\text{estimate} \pm \text{critical value} \times \textbf{standard error of the estimate}$$

Reminder:

Approximate 95% CI can be obtained by $\pm$ twice the standard error for the coefficient.

For confidence interval of regression slope we have a couple of options in R

Option 1.

```
confint(LinearModel.1, 'Body.Mass', level=0.95)

               2.5 %    97.5 %
Body.Mass 0.05565899 0.6689173
```

Option 2.

Goal: Extract the coefficients from the output of the linear model and calculate the approximate SE with nine degrees of freedom. This is the big advantage of saving output from functions as objects. Typically, much more information is about the results are available, and, additionally, can be retrieved for additional use. Extracting coefficients from the objects is the best option, but does come with a learning curve. Let's get started.

First, what information is available in the linear model output beyond the default information? To find out, use the `names()` function

```
names(LinearModel.1)
```

R output:

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

Another way is to use the `summary()` function call.

```
summary(LinearModel.1)$coefficients
              Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -8.4745763  4.6640856 -1.816986 0.10259256
Body.Mass    0.3622881  0.1355472  2.672781 0.02550595
```

How can we get just the standard error for the slope? Note that the estimates are reported in a 2×4 matrix like so:

1,1	1,2	1,3	1,4
2,1	2,2	2,3	2,4

Therefore, to get the standard error for the slope we identify that it is stored in cell `2,2` of the matrix and we write

```
summary(LinearModel.1)$coefficients[2,2]
```

which returns

```
[1] 0.1355472
```

Let's use this information to calculate confidence intervals:

```
slp=summary(LinearModel.1)$coefficients[2,1]
slpErrs=summary(LinearModel.1)$coefficients[2,2]
slp + c(-1,1)*slpErrs*qt(0.975, 9)
```

where `qt()` is the quantile function for the t distribution and "9" is the degrees of freedom from the regression. Results follow.

```
coef + c(-1,1)*errs*qt(0.975, 9)
[1] 0.05565899 0.66891728
```

And for the intercept

```
int=summary(LinearModel.1)$coefficients[1,1]
intErrs=summary(LinearModel.1)$coefficients[1,2]
int + c(-1,1)*intErrs*qt(0.975, 9)
```

Results

```
int + c(-1,1)*intErrs*qt(0.975, 9)
[1] -19.025471   2.076318
```

In conclusion, part of fitting a model includes reporting the estimates of the coefficients (model parameters). And, in general, when estimation is performed, reporting of suitable confidence intervals are expected.

Extract additional statistics from R's linear model function

The `summary()` function is used to report the general results from ANOVA and linear model function output in R software, but additional functions can be used to extract the rest of the output, e.g., coefficient of determination. To complete our example of extracting information from the `summary()` function, we next turn to `summary.lm()` function to see what is available.

At the R prompt type and submit

```
summary.lm(LinearModel.1)
```

This returns the following R output:

```
Call:
lm(formula = Matings ~ Body.Mass, data = bird_matings)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29237 -1.34322 -0.03178  1.33792  2.70763
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.4746     4.6641  -1.817   0.1026
Body.Mass     0.3623     0.1355   2.673   0.0255 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.776 on 9 degrees of freedom
Multiple R-squared:  0.4425,    Adjusted R-squared:  0.3806
F-statistic: 7.144 on 1 and 9 DF,  p-value: 0.02551
```

Looks exactly like the output from `summary()` . Let's look at what is available in the `summary.lm()` function

```
names(summary.lm(LinearModel.1))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

We see some information we got from `summary()` , e.g., "coefficients". If we interrogate the name coefficients like so

```
summary.lm(LinearModel.1)$coefficients
```

we get

```
summary.lm(LinearModel.1)$coefficients
              Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -8.4745763  4.6640856 -1.816986 0.10259256
Body.Mass    0.3622881  0.1355472  2.672781 0.02550595
```

which, again, is a 2×4 matrix (see above)

So to get the standard error for the slope we identify that it is stored in cell `2,2` of the matrix and call it `LinearModel.1$coefficients[2,2]` .

Questions

pending

17.4: OLS, RMA, and smoothing functions

Introduction

OLS or **ordinary least squares** is the most commonly used estimation procedure for fitting a line to the data. For both simple and multiple regression, OLS works by minimizing the sum of the squared residuals. OLS is appropriate when the linear regression **assumptions LINE** apply. In addition, further restrictions apply to OLS including that the predictor variables are fixed and without error. OLS is appropriate when the goal of the analysis is to retrieve a predictive model. OLS describes an asymmetric association between the predictor and the response variable: the slope b_x for $Y \sim b_x X$ will generally not be the same as the slope b_y for $X \sim b_y Y$.

OLS is appropriate for assessing functional relationships (i.e., inference about the coefficients) as long as the assumptions hold. In some literature, OLS is referred to as a **Model I regression**.

Generalized Least Squares

Generalized linear regression is an estimation procedure related to OLS but can be used either when variances are unequal or multicollinearity is present among the error terms.

Weighted Least Squares

A conceptually straightforward extension of OLS can be made to account for situation where the variances in the error terms are not equal. If the variance of $]Y_i$ varies for each X_i, then a weighting function based on the reciprocal of the estimated variance may be used.

$$w_i = \frac{1}{s_i^2}$$

Then, instead of minimizing the squared residuals as in OLS, the regression equation estimates in weighted least squares minimizes the squared residuals summed over the weights.

$$\sum w_i \left(y_i - \hat{y}_i\right)$$

Weighted least squares is a form of generalized least squares. In order to estimate w_i, however, multiple values of Y for each observed X must be available.

Reduced Major Axis

There are many alternative methods available when OLS may not be justified. These approaches, collectively, may be called **Model II regression** methods. These methods are invariably invoked in situations in which both Y and X variables have random error associated with them. In other words, the OLS assumption that the predictor variables are measured without error is violated. Among the more common methods is one called Reduced Major Axis or RMA.

Smoothing functions

Data set: atmospheric carbon dioxide (CO_2) readings Mauna Loa. Source: http://co2now.org/Current-CO2/CO2-Now/noaa-mauna-loa-co2-data.html

Fit curves without applying a known formula. This technique is called **smoothing** and, while there are several versions, the technique involves taking information from groups of observations and using these groups to estimate how the response variable changes with values of the independent variable. Techniques by name include kernel, loess, and spline. Default in the scatter plot command is loess.

CO_2 in parts per million (ppm) plotted by year from 1958 to 2014 the first CO_2 readings were recorded in April 1958; the last data available for this plot was April 2014) (Fig. 17.4.1).

Note:

CO_2 416.71 ppm December 2021, a 0.6% rise since December 2020; https://www.esrl.noaa.gov/gmd/ccgg/trends/

A few words of explanation for Figure 17.4.1. The green line shows the OLS line, and the red line shows the loess smoothing with a smoothing parameter of 0.5 (in `Rcmdr` the slider value reads “5”).

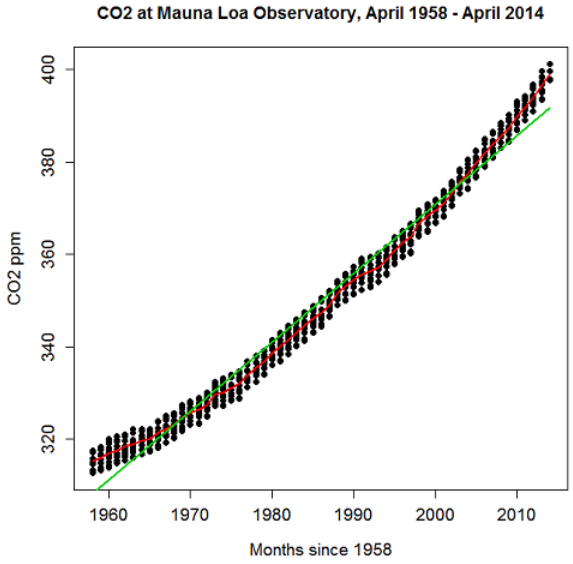


Figure 17.4.1: CO_2 in parts per million (ppm) plotted by year from 1958 to 2014.

R command was started with option settings available in Rcmdr context menu for scatterplot, then additional commands were added

```
scatterplot(CO2~Year, reg.line=lm, grid=FALSE, smooth=TRUE, spread=FALSE, boxplots=FA
span=0.05, lwd=2, xlab="Months since 1958", ylab="CO2 ppm", main="CO2 at Mauna Loa
Observatory, April 1958 - April 2014", cex=1, cex.axis=1.2, cex.lab=1.2, pch=c(16), d
```

The next plot is for ppm CO_2 by month for the year 2013. The plot shows the annual cycle of atmospheric CO_2 in the northern hemisphere.

Again, the smoothing parameter was set to 0.5 and the loess function is plotted in red (Fig. 17.4.2).

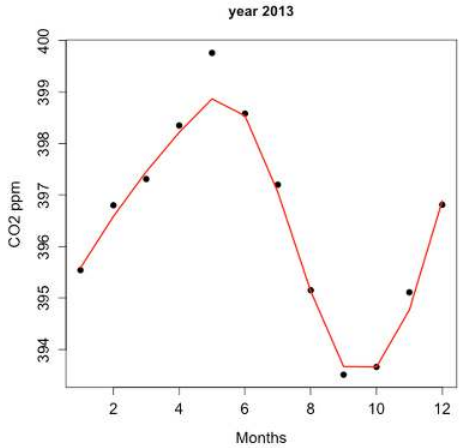


Figure 17.4.2: Plot of ppm CO_2 by month for the year 2013.

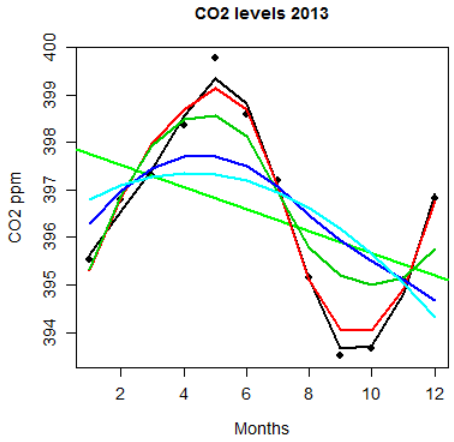
Loess is an acronym short for local regression. Loess is a weighted least squares approach which is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. This percentage of data points is referred to as the **smoothing parameter** and

this parameter may differ for different neighborhoods of points. The idea of loess, in fact, any smoothing algorithm, is to reveal pattern within a noisy sequence of observations. The smoothing parameter can be set to different values, between 0 and 1 is typical.

Note:

Noisy data in this context refers to data comes with random error independent of the true signal, i.e., noisy data has low signal-to-noise ratio. The concept is most familiar in communication.

To get a sense of what the parameter does, Figure 17.4.3 takes the same data as in Figure 17.4.2, but with different values of the smoothing parameter (Fig. 17.4.3).



Parameter	Color
0.5	black
0.75	red
1.0	dark green
2.0	blue
10.0	light blue

Figure 17.4.3: Plot with different smoothing values (0.5 to 10.0).

The R code used to generate the Figure 17.4.3 plot was

```
spanList = c(0.5, 0.75, 1, 2, 10)
reg1 = lm(ppm~Month)
png(filename = "RplotCO2mo.png", width = 400, height = 400, units = "px", pointsize =
plot(Month,ppm, cex=1.2, cex.axis=1.2, cex.lab=1.2, pch=c(16), xlab="Months", ylab="C
abline(reg1,lwd=2,col="green")
for (i in 1:length(spanList))
{
ppm.loess <- loess(ppm~Month, span=spanList[i], Dataset)
ppm.predict <- predict(ppm.loess, Month)
lines(Month,ppm.predict,lwd=2,col=i)
}
```

Note: This is our first introduction to use of a “for” loop.

The CO_2 data constitutes a time series. Instead of loess, a **simple moving average** would be a more natural way to reveal trends. In principle, take a set of nearby points (odd number of points best, keeps the calculation symmetric) and calculate the average. Next, shift the points by a specified time interval (e.g., 7 days), and recalculate the average for the new set of points. See Chapter 20.5 for Time series analysis.

Questions

1. This is a biology class, so I gotta ask: What environmental process explains the shape of the relationship between ppm CO_2 and months of the year as shown in Figure 17.4.2? Hint: NOAA Global Monitoring Laboratory responsible for the CO2 data is located at Mauna Loa Observatory, Hawaii (lat: 19.52291, lon: -155.61586).
2. As I write this question (January 2022), we are 22 months since W.H.O. declared Covid-19 a pandemic (CDC timeline). Omicron variant is now dominant; Daily case counts State of Hawaii from 1 November 2021 to 15 January 2022 reported in data set table.
    1. Make a plot like Figure 17.4.2(days instead of months)
    2. Apply different loess smoothing parameters and re-plot the data. Observe and describe the change to the trend between case reports and days.

Data set

Covid-19 cases reported State of Hawaii from 1 November 2021 to 15 January 2022 (data extracted from Wikipedia)

Date	Cases reported
11/01/21	69
11/02/21	38
11/03/21	176
11/04/21	112
11/05/21	124
11/06/21	97
11/07/21	134
11/08/21	94
11/09/21	79
11/10/21	142
11/11/21	130
11/12/21	138
11/13/21	81
11/14/21	0
11/15/21	146
11/16/21	93
11/17/21	142
11/18/21	226
11/19/21	206
11/20/21	218
11/21/21	107
11/22/21	92

Date	Cases reported
11/23/21	52
11/24/21	115
11/25/21	77
11/26/21	27
11/27/21	135
11/28/21	169
11/29/21	71
11/30/21	79
12/01/21	108
12/02/21	126
12/03/21	125
12/04/21	124
12/05/21	148
12/06/21	90
12/07/21	55
12/08/21	72
12/09/21	143
12/10/21	170
12/11/21	189
12/12/21	215
12/13/21	150
12/14/21	214
12/15/21	282
12/16/21	395
12/17/21	797
12/18/21	707
12/19/21	972
12/20/21	840
12/21/21	707
12/22/21	961
12/23/21	1511
12/24/21	1828
12/25/21	1591
12/26/21	2205

Date	Cases reported
12/27/21	1384
12/28/21	824
12/29/21	1561
12/30/21	3484
12/31/21	3290
01/01/22	2710
01/02/22	3178
01/03/22	3044
01/04/22	1592
01/05/22	2611
01/06/22	4789
01/07/22	3586
01/08/22	4204
01/09/22	4578
01/10/22	3875
01/11/22	2929
01/12/22	3512
01/13/22	3392
01/14/22	3099
01/15/22	5977

17.5: Testing regression coefficients

Introduction

Whether the goal is to create a predictive model or an explanatory model, then there are two related questions the analyst asks about the linear regression model fitted to the data:

1. Does a line actually fit the data
2. Is the linear regression statistically significant?

We will turn to the first question pertaining to fit in time, but for now, focus on the second question.

Like the one-way ANOVA, we have the null and alternate hypothesis for the regression model itself. We write our hypotheses for the regression:

H_O: linear regression fits the data vs. H_A: linear regression does not fit

We see from the output in R and `Rcmdr` that an ANOVA table has been provided. Thus, the test of the regression is analogous to an ANOVA — we partition the overall variability in the response variable into two parts: the first part is the part of the variation that can be explained by there being a linear regression (the linear regression **sum of squares**) plus a second part that accounts for the rest of the variation in the data that is not explained by the regression (the **residual** or **error sum of squares**). Thus, we have

$$SS_{total} = SS_{regression} + SS_{residual}$$

As we did in ANOVA, we calculate **Mean Squares** $MS_x = SS_x / DF_x$, where x refers to either "regression" or "residual" sums of squares and **degrees of freedom**. We then calculate F-values, the test statistics for the regression, to test the null hypothesis.

The degrees of freedom (DF) in simple linear regression are always

$$\begin{aligned} DF_{total} &= n-1 \\ DF_{regresion} &= 1 \\ DF_{residual} &= DF_{total} - DF_{regression} = n-2 \end{aligned}$$

where

$$F = \frac{MS_{regression}}{MS_{residual}}$$

and $F_{DF\,regression}$, $DF_{residual}$ are compared to the critical value at Type I error rate α, with $DF_{regression}$, $DF_{residual}$.

Linear regression inference

Estimation of the slope and intercept is a first step and should be accompanied by the calculation of confidence intervals.

What we need to know if we are to conclude that there's a functional relationship between the X and Y variable is whether the same relationship exists in the population. We've sampled from the population, calculated an equation to describe the relationship between them. However, just as in all cases of inferential statistics, we need to consider the possibility that, through chance alone, we may have committed a Type I error.

The graph below (Fig. 17.5.1) shows a possible outcome under a scenario in which the statistical analyst would likely conclude that there is a statistically significant linear model fit to the data, but the true relationship in the population was a slope of zero. How can this happen? Under this scenario, by chance alone the researchers sampled points (circled in red) from the population that fall along a line. We will conclude that there is linear relationship — that's what are inferential statistics work would indicate — but there was none in the population from which the sampling was done; there would be no way for us to recognize the error except to repeat the experiment — the principal of research **reproducibility** — with a different sample.

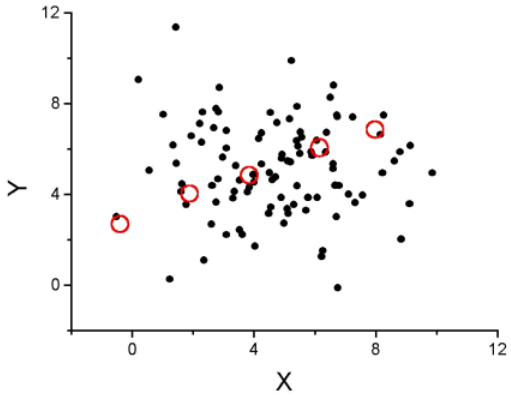


Figure 17.5.1: Scatterplot of hypothetical (x, y) data for which the researcher may obtain a statistically significant linear fit to sample of data from population in which null hypothesis is true relationship between x and y.

So in conclusion, you must keep in mind the meaning of statistical significance in the context of statistical inference: it is inference done on a background of random chance, the chance that sampling from the population leads to a biased sample of subjects.

Note:

If you think about what I did with this data for the Figure 17.5.1 graph, *purposely* selecting data that showed a linear relationship between Y and X (red circles), then you should recognize this as an example of **data dredging** or **p-hacking**, cf. discussion in Head et al (2015); Stefan and Schönbrodt (2023). However, the graph is supposed to be read as if we could do a census and therefore have full knowledge of the true relationship between y and x. The red circles indicate the chance that sampling from a population may sometimes yield incorrect conclusions.

Tests of coefficients

One criterion for a good model is that the coefficients in the model, the intercept and the slope(s) are all statistically significant.

For the statistical of the slope, b_1, we generally treat the test as a two-tailed test of the null hypothesis that the regression slope is equal to zero.

$$H_O : b_1 = 0 \text{ vs. } H_A : b_1 \neq 0$$

Similarly, for the statistical of the intercept, b_0, we generally treat the test as a two-tailed test of the null hypothesis that the Y-intercept is equal to zero.

$$H_O : b_0 = 0 \text{ vs. } H_A : b_0 \neq 0$$

For both slope and intercept we use t-statistics.

$$t = \frac{b}{SE_b}$$

We'll illustrate the tests of the slope and intercept by letting R and `Rcmdr` do the work. You'll find this simple data set at the bottom of this page (scroll or click here). The first variable is the number of `matings`, the second is the size of the paired female, and the third is the size of the paired male. All body mass are in grams.

R code

After loading the worksheet into R and `Rcmdr`, begin by selecting

Rcmdr: Statistics → Fit model → Linear Regression

Note that more than one predictor can be entered, but only one response (dependent) variable may be selected (Fig. 17.5.2

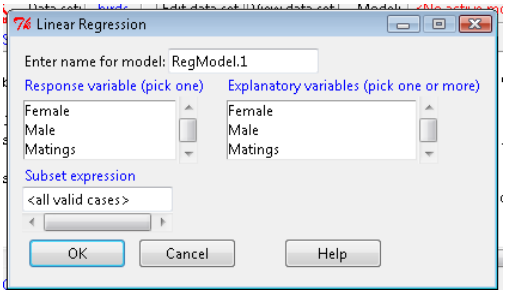


Figure 17.5.2: Screenshot linear regression menu. More than explanatory (predictor or independent) variables may be selected, but only one response (dependent) variable may be selected.

This procedure will handle simple and multiple regression problems. But before we go further, answer these two questions for the data set.

Question 1. What is the Response variable?

Question 2. What is the Explanatory variable?

Answers. See below in the R output to see if you were correct!

If there is only one predictor, then this is a Simple Linear Regression; if more than one predictor is entered, then this is a Multiple Linear Regression. We'll get some more detail, but for now, identify the test of the slope (labeled after the name of the predictor variable), the test of the intercept, and some new stuff.

R output

```
RegModel.1 <- lm=(Matings ~ Female, data=birds)
summary(RegModel.1)
Call: lm(formula = Matings ~ Female, data = birds)
Residuals:
     Min       1Q     Median         3Q        Max
-2.32805 -1.59407   -0.04359    1.77292    2.67195

Coefficients:
             Estimate    Std. Error     t value    Pr(>|t|)
(Intercept)    -8.6175        3.5323      -2.440     0.02528 *
Female          0.3670        0.1042       3.524     0.00243 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.774 on 18 degrees of freedom
Multiple R-squared: 0.4082, Adjusted R-squared: 0.3753 F-statistic: 12.42 on 1 and 18
```

In this example, slope = 0.367 and the intercept = -8.618. The first new term we encounter is called "R-squared" (R^2) — it's also called the **coefficient of determination**. It's the ratio of the sum of squares due to the regression to the total sums of squares. R^2 ranges from zero to 1, with a value of 1 indicating a perfect fit of the regression to the data.

If you are looking for a link between correlation and simple linear regression, then here it is: R^2 is the square of the product-moment correlation, r (see also Chapter 17.2). Thus, $r = \sqrt{R^2}$.

Interpretation of R^2 goes like this: If R^2 is close to zero, then the regression model does not explain much of the variation in the dependent variable; conversely, if R^2 is close to one, then the regression model explains a lot of the variation in the dependent variable.

Did you get the Answers?

Answer 1: Number of `matings`

Answer 2: Size of females

Interpreting the output

Recall that $SS_{total} = SS_{regression} + SS_{residual}$

then $R_2 = SS_{regression}\ SS_{total}$

From the output we see that R^2 was 0.408, which means that about 40% of the variation in numbers of `matings` may be explained by size of the females alone.

To complete the analysis get the ANOVA table for the regression.

Rcmdr: **Models → Hypothesis tests → ANOVA table…**

```
> Anova(RegModel.1, type="II")
Anova Table (Type II tests)
            Sum Sq   Df   F value      Pr(>F)
Female      39.084    1    12.415    0.002427 **
Residuals   56.666   18
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With only one predictor (explanatory) variable, note that the regression test in the ANOVA is the same (has the same probability vs. the null hypothesis) as the test of the slope.

Test two slopes

A more general test of the null hypothesis involving the slope might be

$$H_O : b_1 = b \text{ vs. } H_A : b_1 \neq b$$

where b can be any value, including zero. Again, the t-test would then be used to conduct the test of the slope, where the t-test would have the form

$$t = \frac{b_1 - b_2}{SE_{b_1 - b_2}}$$

where $SE_{b_1 - b_2}$ is the standard error of the difference between the two regression coefficients. We saw something similar to this value back when we did a paired t-test. To obtain $SE_{b_1 - b_2}$, we need the pooled residual mean square and the squared sums of the X values for each of our sample.

First, the pooled (hence the subscript p) residual mean square is calculated as

$$\left(s^2_{y \cdot x}\right)_p = \frac{resSS_1}{resDF_1} + \frac{resSS_2}{resDF_2}$$

where $resSS$ and $resDF$ refer to residual sums of squares and residual degrees of freedom for the first (1) and second (2) regression equations.

Second, the standard error of the difference between regression coefficients (squared!!) is calculated as

$$s_{b_1 - b_2} = \frac{\left(s^2_{y \cdot x}\right)_p}{\sum x_1^1} + \frac{\left(s^2_{y \cdot x}\right)_p}{\sum x_2^2}$$

where the subscript “1” and “2” refer to the X values from the first sample (e.g., the body size values for the males) and the second sample (e.g., the body size values for the females).

Note:

To obtain the squared sum in R and `Rcmdr` , use the Calc function (e.g., to sum the squared X values for the females, use SUM('Female'*'Female')).

We can then use our t-test, with the degrees of freedom now

$$DF = n_1 + n_2 - 4$$

Alternatively, the t-test of two slopes can be written as \[t = \frac{b_{1} - b_{2}}{\left(SE_{b_{1}}^{2} + SE_{b_{2}}^{2}\right)^{2}} \nonumber\]

with again $DF = n_1 + n_2 - 4$.

In this way, we can see a way to test any two slopes for equality. This would be useful if we wanted to compare two samples (e.g., males and females) and wanted to see if the regressions were the same (e.g., metabolic rate covaried with body mass in the same way — that is, the slope of the relationship was the same). This situation arises frequently in biology. For example, we might want to know if male and female birds have different mean field metabolic rates, in which case we might be tempted to use a one-way ANOVA or t-test (since there is one factor with two levels). However, if males and females also differ for body size, then any differences we might see in metabolic rate could be due to differences in metabolic rate or to differences in the covariable of body size. The test is generally referred to as the analysis of covariance (ANCOVA), which is the subject of Chapter 17.6. In brief, ANCOVA allows you to test for mean differences in traits like metabolic rate between two or more groups, but after first accounting for covariation due to another variable (e.g., body size). However, ANCOVA makes the assumption that the relationship between the covariable and the response variable is the same in the two groups. This is the same as saying that the regression slopes are the same. Let's proceed to see how we can compare regression slopes, then move to a more general treatment in Chapter 17.6.

Example

For a sample of 13 tadpoles (*Rana pipiens)*, hatched in the laboratory.

M. Dohm unpublished results from my undergraduate days, You'll find this simple data set at the bottom of this page (scroll or click here).

You should confirm your self, but the slope of the regression equation of oxygen consumption (ml O_2/hour) on body mass (g) was 444.95 (with SE = 65.89). Plot of data shown in Figure 17.5.3

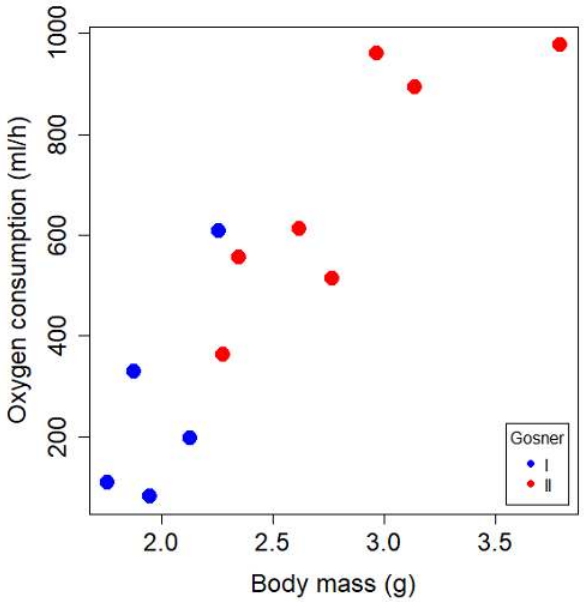


Figure 17.5.3: Scatterplot of oxygen consumption by tadpoles (blue: Gosner developmental stage I; red: Gosner developmental stage II), vs body mass (g).

The project looked at whether metabolism as measured by oxygen consumption was consistent across two developmental stages. Metamorphosis in frogs and other amphibians represents profound reorganization of the organism as the tadpole moves from water

to air. Thus, we would predict some cost as evidenced by change in metabolism associated with later stages of development. Figure 17.5.4 shows a box plot of tadpole oxygen consumption by Gosner (1960) developmental stage.

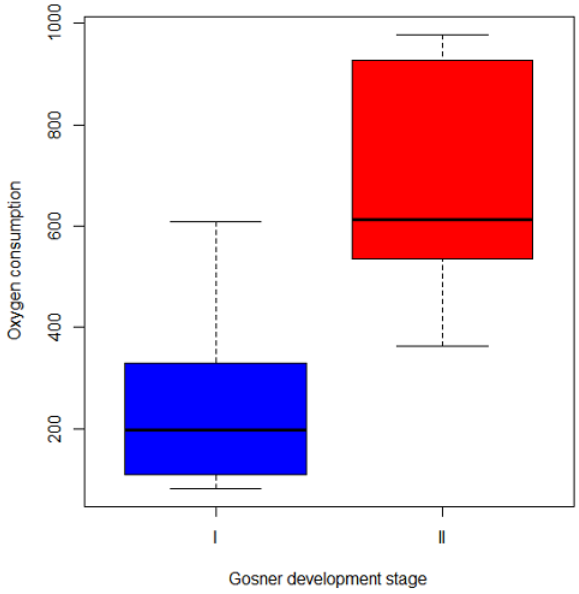


Figure 17.5.4: Boxplot of oxygen consumption by Gosner developmental stages (blue: stage I; red: stage 2).

Looking at Figure 17.5.4 we see a trend consistent with our prediction; developmental stage may be associated with increased metabolism. However, older tadpoles also tend to be larger, and the plot in Figure 17.5.4 does not account for that. Thus, body mass is a **confounding variable** in this example. There are several options for analysis here (e.g., ANCOVA), but one way to view this is to compare the slopes for the two developmental stages. While this test does not compare the means, it does ask a related question: is there evidence of change in rate of oxygen consumption relative to body size between the two developmental stages? The assumption that the slopes are equal is a necessary step for conducting the ANCOVA, which we describe in Chapter 17.6.

So, divide the data set into two groups by developmental stage (12 tadpoles could be assigned to one of two developmental stages; one was at a lower Gosner stage than the others and so is dropped from the subset.

Gosner stage I:

Body mass	VO2
1.76	109.41
1.88	329.06
1.95	82.35
2.13	198
2.26	607.7

Gosner stage II:

Body mass	VO2
2.28	362.71
2.35	556.6
2.62	612.93
2.77	514.02
2.97	961.01

Body mass	VO2
3.14	892.41
3.79	976.97

The slopes and standard errors were

	Gosner Stage I	Gosner stage II
slope	750.0	399.9
standard error of slope	444.6	111.2

Rcmdr: Models – Compare model coefficients..

```
compareCoefs(gosI.1, gosII.1)
Calls:
1: lm(formula = VO2 ~ Body.mass, data = gosI)
2: lm(formula = VO2 ~ Body.mass, data = gosII)

            Model 1    Model 2
(Intercept)   -1232       -441
SE              891        321

Body.mass       750        400
SE              445        111
```

Are the two slopes equal?

Looking at the table, we would say No, because the slopes look different (750 vs 399.9). However, the errors are large and, given this is a small data set, we need to test statistically; are the slopes indistinguishable $(H_O : b_I = b_{II})$, where b_I is the slope for the Gosner Stage I subset and b_{II} is the slope for the Gosner Stage II subset?

To use our tests discussed above, we need the sum of squared X values for Gosner Stage I and sum of squared X values for Gosner stage II results. We can get these from the ANOVA tables. Recall that we can apply:

ANOVA regression Gosner Stage I

```
Anova(gosI.1, type="II")
Anova Table (Type II tests)

Response: VO2
            Sum Sq   Df    F value   Pr(>F)
Body.mass    89385    1     2.8461   0.1902
Residuals    94220    3
```

ANOVA regression Gosner Stage II

```
Anova(gosII.1, type="II")
Anova Table (Type II tests)

Response: VO2
```

```
             Sum Sq    Df   F value  Pr(>F)
Body.mass    258398     1    12.935  0.0156 *
Residuals     99882     5
```

$SS_{gosnerI}$ = 89385

$SS_{gosnerII}$ = 258398

We also need the residual Mean Squares ($SS_{residual}/DF_{residual}$) from the ANOVA tables

$MS_{gosnerI}$ = 94220/3 = 31406.67

$MS_{gosnerII}$ = 99882/5 = 19976.4

Therefore, the pooled residual MS is $\left(s^2_{\bar{x}\cdot y}\right)_p = 51383.07$ and the pooled SE of the difference is $s_{b_1-b_2} = 3440.359$ using the formulas above.

Now, we plug in the values to get a t-test: $t = \dfrac{750-444.6}{3440.359} = 0.0887698$.

The DF for this t-test are $n_1+n_2-4 = 4+6-4 = 6$.

Using Table of Student's t distribution (Appendix), I find the two-tailed critical value for t at alpha = 5% with DF = 6 is equal to 3.758. Since $0.0887698 << 3.758$ we cannot conclude that the two slopes are statistically different.

Questions

Metabolic rates like oxygen consumption over time are well-known examples of allometric relationships. That is, many biological variables (e.g., $\dot{V}O_2$ is related as $a \cdot M^b$, where M is body mass, b is scaling exponent (the slope!), and a is a constant (the intercept!)) are best evaluated on log-log scale. Redo the linear regression of oxygen consumption vs. body mass for the tadpoles, but this time, apply log10-transform to `VO2` and to `Body.mass` .

Data in this page, bird matings

Body mass	Matings
29	0
29	2
29	4
32	4
32	2
35	6
36	3
38	3
38	5
38	8
40	6

Data in this page, Oxygen consumption, $\dot{V}O_2$, of Anuran tadpoles

Gosner	Body mass	VO2
NA	1.46	170.91
I	1.76	109.41

I	1.88	329.06
I	1.95	82.35
I	2.13	198
II	2.28	362.71
I	2.26	607.7
II	2.35	556.6
II	2.62	612.93
II	2.77	514.02
II	2.97	961.01
II	3.14	892.41
II	3.79	976.97

Gosner refers to Gosner (1960), who developed a criteria for judging metamorphosis staging.

17.6: ANCOVA - analysis of covariance

Introduction

Analysis of covariance (ANCOVA) is intended to help with analysis of designs with categorical treatment variables on some response (dependent) variable, but a known **confounding variable** is also present. Thus, the researcher is also likely to know of additional ratio scale variables that **covary** with the response variable and, moreover, must be included in the experimental design in some way.

Take for example the well-known relationship between body size and whole-animal metabolic rate as measured by rates of carbon dioxide production or rates of oxygen consumption for aerobic organisms. We may be interested in how addition or blocking of stress hormones affects resting metabolism; we may be interested in comparing men and women for activity metabolism, and so on. We'd like to know if the regressions were the same (e.g., metabolic rate covaried with body mass in the same way — that is, the slope of the relationship was the same).

This situation arises frequently in biology. For example, we might want to know if male and female birds have different mean field metabolic rates, in which case we might be tempted to use a one-way ANOVA or t-test (since one factor with two levels). However, if males and females also differ for body size, then any differences we might see in metabolic rate could be due to differences in metabolic rate are confounded by differences in the covariable body size. We already discussed one approach to correction: calculate a ratio. Thus, a logical approach to correcting or **normalizing** for the covariation would be to divide body mass (units of kilograms) into metabolic rate (e.g., volume of oxygen, O_2, consumed), and make comparisons, say, among different species, on mass-specific trait $\left(\frac{ml\ O_2}{hours \cdot mass}\right)$. However, because the regression between mass and metabolic rate is allometric, i.e., not equal to one, the ratio does not, in fact normalize for body mass. We made this point in Chapter 6.2, and remarked that analysis of covariance ANCOVA was a solution.

ANCOVA allows you to test for mean differences in traits like metabolic rate between two or more groups, but only after first accounting for covariation due to another variable (e.g., body size). However, ANCOVA makes the assumption that relationship between the covariable and the response variable is the same in the two groups. This is the same as saying that the regression slopes are the same. We discussed how to use **t-test** to test hypothesis of **equal slopes** between regression models in Chapter 17.5, but a more elegant way is to include this in your model.

Example

We return to our sample of 13 tadpoles (*Rana pipiens)*, hatched in the laboratory. I've repeated the data set in this page, scroll or click here.

Our linear model was `VO2 ~ Body.mass` and a scatterplot of the data set is shown in Figure 17.6.1 (a repeat of Figure 17.5.3 from Chapter 17.5, but now points identified to developmental group).

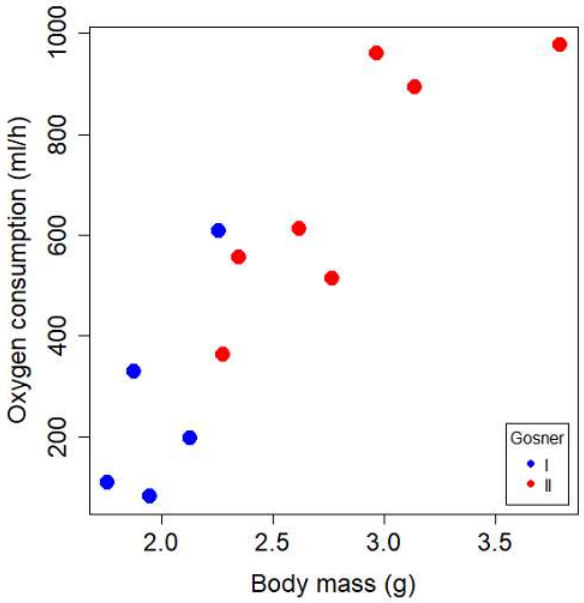


Figure 17.6.1: Copy and Paste Caption here. (Copyright; author via source)

The project looked at whether metabolism as measured by oxygen consumption was consistent across two developmental stages. Metamorphosis in frogs and other amphibians represents profound reorganization of the organism as the tadpole moves from water to air. Thus, we would predict some cost as evidenced by change in metabolism associated with later stages of development. Figure 17.6.2 shows a box plot of tadpole oxygen consumption by Gosner (1960) developmental stage (Figure 17.6.2 is a repeat of Figure 17.5.4 from Chapter 17.5).

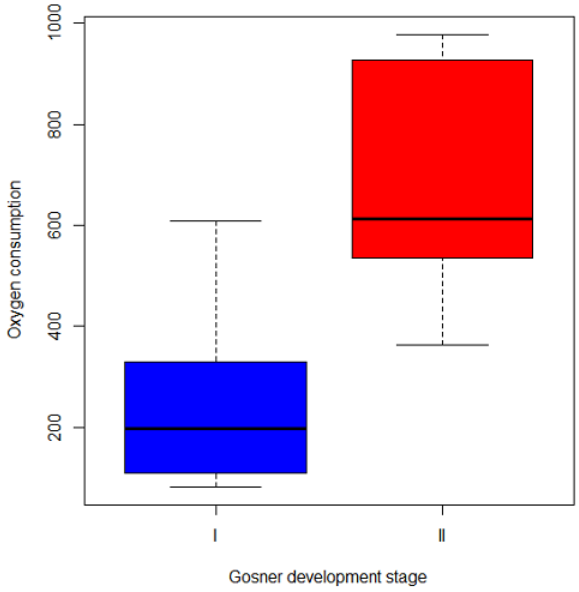


Figure 17.6.2: Boxplot of oxygen consumption by Gosner developmental stages.

Looking at Figure 17.6.2 we see a trend consistent with our prediction; developmental stage may be associated with increased metabolism. However, older tadpoles also tend to be larger, and the plot in Figure 17.6.2 does not account for that. Thus, body mass is a **confounding variable** in this example. There are several options for analysis here (e.g., ANCOVA), but one way to view this is to compare the slopes for the two developmental stages. While this test does not compare the means, it does ask a related question: is there evidence of change in rate of oxygen consumption relative to body size between the two developmental stages? The assumption that the slopes are equal is a necessary step for conducting the ANCOVA.

So, divide the data set into two groups by developmental stage (12 tadpoles could be assigned to one of two developmental stages; one was at a lower Gosner stage than the others and so is dropped from the subset.

Gosner stage I

Body mass	VO2
1.76	109.41
1.88	329.06
1.95	82.35
2.13	198
2.26	607.7

Gosner stage II

Body mass	VO2
2.28	362.71
2.35	556.6
2.62	612.93
2.77	514.02
2.97	961.01
3.14	892.41
3.79	976.97

The slopes and standard errors we obtained in Chapter 17.5 were

	Gosner Stage I	Gosner stage II
slope	750.0	399.9
standard error of slope	444.6	111.2

Make a plot (Figure 17.6.3).

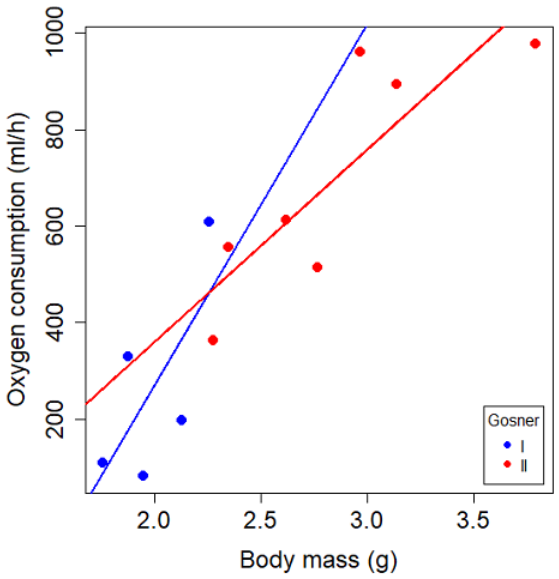


Figure 17.6.3: Scatterplot with best-fit regression lines of `VO2` by `Body.mass` for Gosner State I (in blue) and Gosner Stage II (in red) tadpoles.

R code for plot in Figure 17.6.3

```
#Used Rcmdr scatterplot(), then modified code
scatterplot(VO2~Body.mass | Gosner, regLine=FALSE, smooth=FALSE,
boxplots=FALSE, xlab="Body mass (g)", ylab="Oxygen consumption (ml/h)",
main="", cex=1.4, cex.axis=1.5, cex.lab=1.5, pch=c(19,19), by.groups=TRUE,
col=c("blue","red"), grid=FALSE,
legend=list(coords="bottomright"), data=Tadpoles)
#Get regression equations for groups, subset by Gosner
abline(lm(VO2~Body.mass, data=Stage01), lty=1, lwd=2, col="blue")
abline(lm(VO2~Body.mass, data=Stage02), lty=1, lwd=2, col="red")
```

Returning to the important question, are the two slopes statistically indistinguishable $(H_O : b_I = b_{II})$, where b_I is the slope for the Gosner Stage I subset and b_{II} is the slope for the Gosner Stage II subset? We look at the plot, and since the lines cross, we tend to see a difference. Of course, we need to consider that our perception of slope differences may simply be chance, especially because the sample size is small. Proceed to test.

R code

The ANCOVA is a new ANOVA model where the factor variables are adjusted or corrected for the effects of the continuous variable.

R code for ANCOVA example, **crossed** or **interaction** model.

```
tadpole.1 <- lm(VO2 ~ Body.mass*Gosner, data=example.Tadpole)
summary(tadpole.1)
Anova(tadpole.1, type="II")
```

Output:

```
summary(tadpole.1)

Call:
```

```
lm(formula = VO2 ~ Body.mass * Gosner, data = example.Tadpole)

Residuals:
    Min      1Q  Median      3Q     Max
-167.80 -117.93   13.81   94.66  214.65

Coefficients:
                      Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)            -1231.6       783.0    -1.573   0.1544
Body.mass                750.0       390.7     1.919   0.0912 .
Gosner[T.II]             790.4       859.2     0.920   0.3845
Body.mass:Gosner[T.II]  -350.1       409.5    -0.855   0.4174
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 155.8 on 8 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.821, Adjusted R-squared: 0.7539
F-statistic: 12.23 on 3 and 8 DF, p-value: 0.002336
```

This provides the coefficients for the first factor (GII) and then the differences in the coefficient for the second factor. You can just add the second coefficient to the first so they're on the same scale.

```
Anova Table (Type II tests)

Response: VO2
                  Sum Sq   Df    F value     Pr(>F)
Body.mass         330046    1    13.6030   0.006146 **
Gosner              5630    1     0.2321   0.642908
Body.mass:Gosner   17736    1     0.7310   0.417423
Residuals 194102 8
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Suggests interaction is not significant, i.e., the slopes are not different.

We can then proceed to check to see if the intercepts are different, now that we've confirmed no significant difference in slope.

R code for ANCOVA as **additive** model

```
tadpole.2 <- lm(VO2 ~ Body.mass + Gosner, data=example.Tadpole)
summary(tadpole.2)
Anova(tadpole.2, type="II")
```

Output:

```
> summary(tadpole.2)

Call:
```

```
lm(formula = VO2 ~ Body.mass + Gosner, data = example.Tadpole)

Residuals:
    Min       1Q    Median      3Q      Max
-163.12  -125.53    -20.27   83.71   228.56

Coefficients:
                  Estimate   Std. Error     t value    Pr(>|t|)
(Intercept)         -595.37       239.87      -2.482     0.03487 *
Body.mass            431.20       115.15       3.745     0.00459 **
Gosner[T.II]          64.96       132.83       0.489     0.63648
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 153.4 on 9 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.8047, Adjusted R-squared: 0.7613
F-statistic: 18.54 on 2 and 9 DF, p-value: 0.0006432

Anova(tadpole.2, type="II")
Anova Table (Type II tests)

Response: VO2
             Sum Sq    Df    F value      Pr(>F)
Body.mass     330046     1    14.0221   0.004593 **
Gosner          5630     1     0.2392   0.636482
Residuals     211839     9
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that there is no test of interaction in the added model. This model would be appropriate IF the slopes are equal.

Instead of the additive model, try a **nested** model, with body mass nested within stage.

```
tadpole.3 <- lm(VO2 ~ Body.mass/Gosner, data=example.Tadpole)

summary(tadpole.3)

Call:
lm(formula = VO2 ~ Body.mass/Gosner, data = example.Tadpole)

Residuals:
    Min       1Q    Median      3Q      Max
-168.66  -131.14    -20.28   90.33   225.36

Coefficients:
                Estimate      Std. Error    t value Pr(>|t|)
(Intercept)           -575.10      319.51     -1.800   0.1054
```

```
Body.mass               423.65      162.50      2.607   0.0284 *
Body.mass:Gosner[T.II]  21.95        63.73      0.344   0.7384
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 154.4 on 9 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.8021, Adjusted R-squared: 0.7581
F-statistic: 18.24 on 2 and 9 DF, p-value: 0.0006823
```

This gets the true coefficient (nested `lm()` version).

The two test different hypotheses:

`lm(VO2 ~ Body.mass * Gosner)` tests whether or not the regression has a nonzero slope.

`lm(VO2 ~ Body.mass */Gosner)` test whether or not the slopes and intercepts from different factors are statistically significant.

Questions

1. An OLS approach was used for the analysis of tadpole oxygen consumption body mass. Consider the RMA approach — would that be a more appropriate regression model? Explain why or why not.
2. Consider an experiment in which you plan to administer a treatment that has a carry-over effect. For example, Compare and contrast “crossed” and “nested” designs.
3. True or False. The nested design option for the ANCOVA assumes the slopes for the two groups of tadpoles for the regression line of `VO2` by `Body.mass` are equal. Explain your choice.
4. Metabolic rates like oxygen consumption over time are well-known examples of allometric relationships. That is, many biological variables (e.g., VO2 is related as aM^b, where M is body mass, slope b is scaling exponent), and best evaluated on log-log scale. Repeat the analysis above on $\log_{10}$-transformed `VO2` and `Body.mass` for
   - crossed design (e.g., tadpole.1 model)
   - added design (e.g., tadpole.2 model)
   - nested design (e.g., tadpole.3 model)
5. Create the plot and add the fitted lines from crossed design to the plot.

Note:

About log-transform of a variable. The most straight-forward tact is to create two new variables. For example,

```
lgVO2 <- log(VO2)
```

Another option is to transform the variables within the call to `lm()` function. For example, try

```
lm(log(VO2) ~ log(Body.mass ), data=example.Tadpole)
```

Hint: don’t forget to attach your data set to avoid having to call the variable as, for example, `example.Tadpole$VO2`

Data sets

Oxygen consumption, $dotVO_2$, of Anuran tadpoles, dataset= `example.Tadpole`

Gosner	Body mass	VO2
NA	1.46	170.91

I	1.76	109.41
I	1.88	329.06
I	1.95	82.35
I	2.13	198
II	2.28	362.71
I	2.26	607.7
II	2.35	556.6
II	2.62	612.93
II	2.77	514.02
II	2.97	961.01
II	3.14	892.41
II	3.79	976.97

Gosner refers to Gosner (1960), who developed a criteria for judging metamorphosis staging.

17.7: Regression model fit

Introduction

In Chapter 17.5 and 17.6 we introduced the example of tadpoles body size and oxygen consumption. We ran a simple linear regression, with the following output from R

```
RegModel.1 <- lm(VO2~Body.mass, data=example.Tadpole)

summary(RegModel.1)

Call:
lm(formula = VO2 ~ Body.mass, data = example.Tadpole)

Residuals:
    Min      1Q  Median      3Q     Max
-202.26 -126.35   30.20   94.01  222.55

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)   -583.05       163.97   -3.556   0.00451 **
Body.mass      444.95        65.89    6.753 0.0000314 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 145.3 on 11 degrees of freedom
Multiple R-squared: 0.8057, Adjusted R-squared: 0.788
F-statistic: 45.61 on 1 and 11 DF, p-value: 0.00003144
```

You should be able to pick out the estimates of slope and intercept from the table (intercept was -583 and slope was 445). Additionally, as part of your interpretation of the model, you should be able to report how much variation in VO2 was explained by tadpole body mass (**coefficient of determination**, R^2, was 0.81, which means about 81% of variation in oxygen consumption by tadpoles is explained by knowing the body mass of the tadpole.

What's left to do? We need to evaluate how well our model fits the data, i.e., we evaluate regression model fit. This we can do by evaluating the error components relative to the portion of the model that explains the data. Additionally, we can perform a number of diagnostics of the model relative to the assumptions we made to perform linear regression. These diagnostics form the subject of Chapter 17.8. Here, we ask how well does the model

$$\dot{V}O_2 = b_0 + b_1(Bodymass)$$

fit the data?

Model fit statistics

The second part of fitting a model is to report how well the model fits the data. The next sections apply to this aspect of model fitting. The first area to focus on is the magnitude of the residuals: the greater the spread of residuals, the less well a fitted line explains the data.

In addition to the output from `lm()` function, which focuses on the coefficients, we typically generate the ANOVA table also.

```
Anova(RegModel.1, type="II")
Anova Table (Type II tests)
```

```
Response: VO2
               Sum Sq   Df    F value       Pr(>F)
Body.mass      962870    1     45.605   0.00003144 ***
Residuals      232245   11
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Standard error of regression

S, the **Residual Standard Error** (aka **Standard error of regression**), is an overall measure to indicate the accuracy of the fitted line: it tells us how good the regression is in predicting the dependence of response variable on the independent variable. A large value for S indicates a poor fit. One equation for **S** is given by

$$\mathbf{S} = \sqrt{\frac{SS_{residual}}{n-2}}$$

In the above example, $\mathbf{S} = 145.3$ (underlined, bold in regression output above). We can see how if $SS_{residual}$ is large, **S** will be large indicating poor fit of the linear model to the data. However, by itself **S** is not of much value as a diagnostic as it is difficult to know what to make of 145.3, for example. Is this a large value for **S** ? Is it small? We don't have any context to judge **S**, so additional diagnostics have been developed.

Coefficient of determination

R^2, the **coefficient of determination**, is also used to describe model fit. R^2, the square of the simple product moment correlation r, can take on values between 0 and 1 (0% to 100%). A good model fit has a high R^2 value. In our example above, $R^2 = 0.8057$ or 80.57%. One equation for R^2 is given by

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

A value of R^2 close to 1 means that the regression "explains" nearly all of the variation in the response variable, and would indicate the model is a good fit to the data. Note that the coefficient of determination, R^2, is the squared value of r, the product moment correlation.

Adjusted R-squared

Before moving on we need to remark on the difference between R^2 and adjusted R^2. For Simple Linear Regression there is but one predictor variable, X; for multiple regression there can be many additional predictor variables. Without some correction, R^2 will increase with each additional predictor variables. This doesn't mean the model is more useful, however, and in particular, one cannot compare R^2 between models with different numbers of predictors. Therefore, an adjustment is used so that the coefficient of determination remains a useful way to assess how reliable a model is and to permit comparisons of models. Thus, we have the Adjusted $\bar{R}^2$, which is calculated as

$$\bar{R}^2 = 1 - \frac{SS_{residual}}{SS_{total}} \cdot \frac{DF_{total}}{DF_{residual}}$$

In our example above, Adjusted $R^2 = 0.3806$ or 38.06%.

Which should you report? Adjusted R^2, because it is independent of the number of parameters in the model.

Both $\bar{R}^2$ and **S** are useful for regression diagnostics, a topic which we will discuss next (Chapter 17.8).

Questions

1. True or False. The simple linear regression is called a "best fit" line because it maximizes the squared deviations for the difference between observed and predicted Y values.
2. True or False. Residuals in regression analysis are best viewed as errors committed by the researcher. If the experiment was designed better, or if the instrument was properly calibrated, then residuals would be reduced. Explain your choice.

3. The USA is finishing the 2020 census as I write this note. As you know, the census is used to reapportion Congress and also to determine the number of electoral college votes. In honor of the election for US President that's just days away, in the next series of questions in this Chapter and subsequent sections of Chapter 17 and 18, I'll ask you to conduct a regression analysis on the electoral college. For starters, make the regression of Electoral votes on the 2010 census population. (Ignore for now the other columns, just focus on POP_2019 and Electoral.) Report the
    - regression coefficients (slope, intercept)
    - percent of the variation in electoral college votes explained by the regression $\left(R^2\right)$.
4. Make a scatterplot and add the regression line to the plot

Data set

State	Region	Division	POP_2010	POP_2019	Electoral
Alabama	South	East South Central	4779736	4903185	9
Alaska	West	Pacific	710231	731545	3
Arizona	West	Mountain	6392017	7278717	11
Arkansas	South	West South Central	2915918	3017804	6
California	West	Pacific	37253956	39512223	55
Colorado	West	Mountain	5029196	5758736	9
Connecticut	Northeast	New England	3574097	3565287	7
Delaware	South	South Atlantic	897934	982895	3
District of Columbia	South	South Atlantic	601723	705749	3
Florida	South	South Atlantic	18801310	21477737	29
Georgia	South	South Atlantic	9687653	10617423	16
Hawaii	West	Pacific	1360301	1415872	4
Idaho	West	Mountain	1567582	1787065	4
Illinois	Midwest	East North Central	12830632	12671821	20
Indiana	Midwest	East North Central	6483802	6732219	11
Iowa	Midwest	West North Central	3046355	3155070	6
Kansas	Midwest	West North Central	2853118	2913314	6
Kentucky	South	East South Central	4339367	4467673	8
Louisiana	South	West South Central	4533372	4648794	8
Maine	Northeast	New England	1328361	1344212	4
Maryland	South	South Atlantic	5773552	6045680	10
Massachusetts	Northeast	New England	6547629	6892503	11
Michigan	Midwest	East North Central	9883640	9883635	16
Minnesota	Midwest	West North Central	5303925	5639632	10
Mississippi	South	East South Central	2967297	2976149	6
Missouri	Midwest	West North Central	5988927	6137428	10
Montana	West	Mountain	989415	1068778	3

State	Region	Division	POP_2010	POP_2019	Electoral
Nebraska	Midwest	West North Central	1826341	1934408	5
Nevada	West	Mountain	2700551	3080156	6
New Hampshire	Northeast	New England	1316470	1359711	4
New Jersey	Northeast	Mid-Atlantic	8791894	8882190	14
New Mexico	West	Mountain	2059179	2096829	5
New York	Northeast	Mid-Atlantic	19378102	19453561	29
North Carolina	South	South Atlantic	9535483	10488084	15
North Dakota	Midwest	West North Central	672591	762062	3
Ohio	Midwest	East North Central	11536504	11689100	18
Oklahoma	South	West South Central	3751351	3956971	7
Oregon	West	Pacific	3831074	4217737	7
Pennsylvania	Northeast	Mid-Atlantic	12702379	12801989	20
Rhode Island	Northeast	New-England	1052567	1059361	4
South Carolina	South	South-Atlantic	4625364	5148714	9
South Dakota	Midwest	West-North-Central	814180	884659	3
Tennessee	South	East-South-Central	6346105	6829174	11
Texas	South	West-South-Central	25145561	28995881	38
Utah	West	Mountain	2763885	3205958	6
Vermont	Northeast	New-England	625741	623989	3
Virginia	South	South-Atlantic	8001024	8535519	13
Washington	West	Pacific	6724540	7614893	12
West Virginia	South	South-Atlantic	1852994	1792147	5
Wisconsin	Midwest	East-North-Central	5686986	5822434	10
Wyoming	West	Mountain	563626	578759	3

17.8: Assumptions and model diagnostics for simple linear regression

Introduction

The assumptions for all linear regression:

1. **Linear** model is appropriate.
   The data are well described (fit) by a linear model.
2. **Independent** values of Y and equal variances.
   Although there can be more than one Y for any value of X, the Y's cannot be related to each other (that's what we mean by independent). Since we allow for multiple Y's for each X, then we assume that the variances of the range of Y's are equal for each X value (this is similar to our ANOVA assumptions for equal variance by groups). Another term for equal variances is **homoscedasticity**.
3. **Normality**.
   For each X value there is a normal distribution of Y's (think of doing the experiment over and over).
4. **Error**
   The residuals (error) are normally distributed with a mean of zero.

Note the mnemonic device: **Linear, Independent, Normal, Error** or **LINE**.

Each of the four elements will be discussed below in the context of **Model Diagnostics.** These assumptions apply to how the model fits the data. There are other assumptions that, if violated, imply you should use a different method for estimating the parameters of the model.

Ordinary least squares makes the additional assumption about the quality of the independent variable that e that measurement of X is done without error. Measurement error is a fact of life in science, but the influence of error on regression differs if the error is associated with the dependent or independent variable. Measurement error in the dependent variable increases the **dispersion of the residuals** but will not affect the estimates of the coefficients; error associated with the independent variables, however, will affect estimates of the slope. In short, error in X leads to **biased estimates** of the slope.

The equivalent, but less restrictive practical application of this assumption is that the error in X is at least negligible compared to the measurements in the dependent variable.

Multiple regression makes one more assumption, about the relationship between the predictor variables (the X variables). The assumption is that there is no multicollinearity, a subject we will bring up next time (see Chapter 18).

Model diagnostics

We just reviewed how to evaluate the estimates of the coefficients of the model. Now we need to address a deeper meaning — how well the model explains the data. Consider a simple linear regression first. If $H_O : b = 0$ is not rejected, then the slope of the regression equation is taken to not differ from zero. We would conclude that if repeated samples were drawn from the population, on average, the regression equation would not fit the data well (lots of scatter) and it would not yield useful prediction.

However, recall that we assume that the fit is linear. One assumption we make in regression is that a line can, in fact, be used to describe the relationship between X and Y.

Here are two very different situations where the slope = 0.

> Example 1. Linear Slope = 0, no relationship between X and Y
>
> Example 2. Linear Slope = 0, a significant relationship between X and Y

But even if $H_O : b = 0$ is rejected (and we conclude that a linear relationship between X and Y is present), we still need to be concerned about the fit of the line to the data — the relationship may be more nonlinear than linear, for example. Here are two very different situations where the slope is not equal to 0.

> Example 3. Linear Slope > 0, a linear relationship between X and Y
>
> Example 4. Linear Slope > 0, curve-linear relationship between X and Y

How can you tell the difference? There are many **regression diagnostic tests**, many more than we can cover, but you can start with looking at the **coefficient of determination** (low R^2 means low fit to the line), and we can look at the pattern of residuals plotted

against the either the predicted values or the X variables (my favorite). The important points are:

1. In linear regression, you fit a model (the slope + intercept) to the data;
2. We want the usual hypothesis tests (are the coefficients different from zero?) and
3. We need to check to see if the model fits the data well. Just like in our discussions of chi-square, a "perfect fit would mean that the difference between our model and the data would be zero.

Graph options

Using residual plots to diagnose regression equations

Yes, we need to test the coefficients (intercept $H_O = 0$; slope $H_O = 0$) of a regression equation, but we also must decide if a regression is an appropriate description of the data. This topic includes the use of **diagnostic tests** in regression. We address this question chiefly by looking at

1. **scatterplots** of the independent (predictor) variable(s) vs. dependent (response) variable(s).
   what patterns appear between X and Y? Do your eyes tell you "Line"? "Curve"? "No relation"?
2. **coefficient of determination**
   closer to zero than to one?
3. **patterns of residuals** plotted against the X variables (other types of residual plots are used to, this is one of my favorites)

Our approach is to utilize graphics along with statistical tests designed to address the assumptions.

One typical choice is to see if there are patterns in the residual values plotted against the predictor variable. If the LINE assumptions hold for your data set, then the residuals should have a mean of zero with scatter about the mean. Deviations from LINE assumptions will show up in residual plots.

Here are examples of POSSIBLE outcomes:

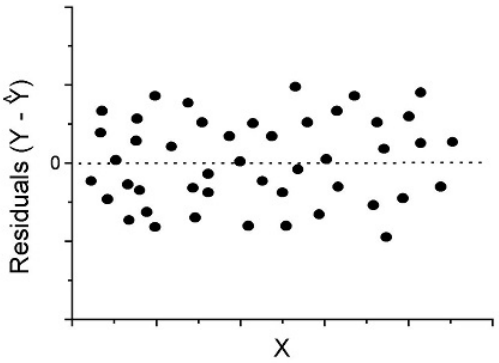


Figure 17.8.1: An ideal plot of residuals.

Solution: Proceed! Assumptions of linear regression met.

Compare to plots of residuals that differ from the ideal.

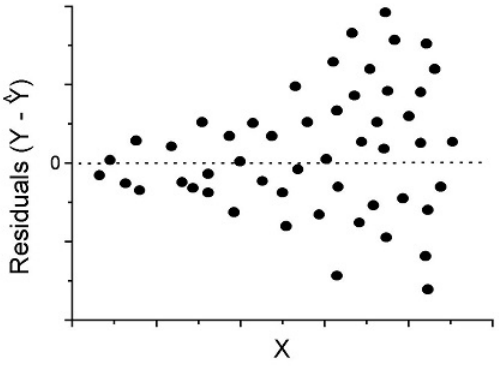


Figure 17.8.2: We have a problem. Residual plot shows **unequal variance** (aka **heteroscedasticity**).

Solution. Try a transform like the $\log_{10}$-transform.

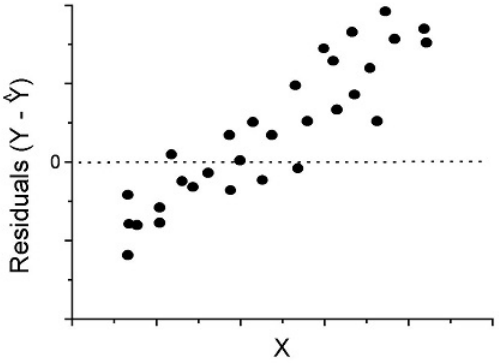


Figure 17.8.3: Problem. Residual plot shows **systematic trend**.

Solution. Linear model a poor fit; may be related to measurement errors for one or more predictor variables. Try adding an additional predictor variable or model the error in your general linear model.

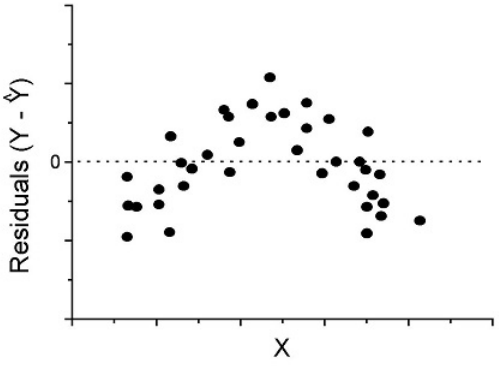


Figure 17.8.4: Problem. Residual plot shows **nonlinear trend**.

Solution. Transform data or use more complex model.

This is a good time to mention that in statistical analyses, one often needs to do multiple rounds of analyses, involving description and plots, tests of assumptions, tests of inference. With regression, in particular, we also need to decide if our model (e.g., linear equation) is a good description of the data.

Diagnostic plot examples

Return to our `example.Tadpole` dataset. To obtain residual plots, **Rcmdr: Models → Graphs → Basic diagnostic plots** yields four graphs.

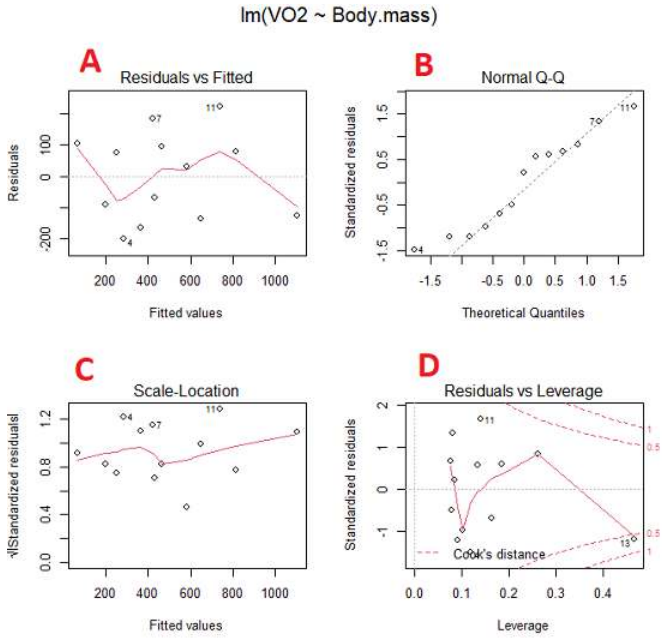


Figure 17.8.5: Basic diagnostic plots. A: residual plot; B: Q-Q plot of residuals; C: Scale-location (aka spread-location) plot; D: leverage residual plot.

In brief, we look at the plots:

A, the **residual plot**, to see if there are trends in the residuals. We are looking for a spread of points equally above and below the mean of zero. In Figure 17.8.5 we count seven points above and six points below zero so there's no indication of a trend in the residuals vs the fitted `VO2` (Y) values.

B, the **Q-Q plot** is used to see if normality holds. As discussed before, if our data are more or less normally distributed, then points will fall along a straight line in a Q-Q plot.

C, the **Scale-** or **spread-location plot** is used to verify equal variances of errors.

D, **Leverage plot** — looks to see if an outlier has leverage on the fit of the line to the data, i.e., changes the slope. Additionally, provides location of **Cook's distance** measure (dashed red lines). Cook's distance measures the effect on the regression by removing one point at a time and then fitting a line to the data. Points outside the dashed lines have influence.

Note:

A note of caution about over-thinking with these plots. R provides a red line to track the points. However, these lines are guides, not judges. We humans are generally good at detecting patterns, but with data visualization, there is the risk of seeing patterns where none exits. In particular, recognizing randomness is not easy. If anything, we may tend to see patterns where none exist, termed apophenia. So yes, by all means look at the graphs, but do so with a plan: red line more or less horizontal? Then there is no pattern and the regression model is a good fit to the data.

Statistical test options

After building linear models, run statistical diagnostic tests that compliment graphics approaches. These are available via

Rcmdr: Models → Numerical diagnostics

Variance inflation factors (VIF): used to detect multicollinearity among the predictor variables. If correlations are present among the predictor variables, then you can't rely on the the coefficient estimates — whether predictor A causes change in the response variable depends on whether the correlated B predictor is also included in the model. If correlation between predictor A and B, the statistical effect is increased variance associated with the error of the coefficient estimates. There are

VIF for each predictor variable. A VIF of one means there is no correlation between that predictor and the other predictor variables. A VIF of 10 is taken as evidence of serious multicollinearity in the model.

Breusch-Pagan test for heteroscedasticity… Recall that heteroscedasticity is another name for unequal variances. The test statistic can be calculated as $\chi^2 \sim nR^2$

Durbin-Watson for autocorrelation

RESET test for nonlinearity

Questions

1. Referring to Figures $17.8.1$– $17.8.4$on this page, which plot best suggests a regression line fits the data?
2. Return to the electoral college data set and your linear models of Electoral vs. POP_2010 and POP_2019. Obtain the four basic diagnostic plots and comment on the fit of the regression line to the electoral college data.
   - Residual plot
   - Q-Q plot
   - Scale-location plot
   - Leverage plot
3. With respect to your answers in question 2, how well does the electoral college system reflect the principle of one person, one vote?

CHAPTER OVERVIEW

18: Multiple Linear Regression

Introduction

This is the second part of our discussion about general linear models. In this chapter we extend linear regression from one (Chapter 17) to many predictor variables. We also introduce **logistic regression**, which uses logistic function to model **binary outcome variables**. Extensions to address **ordinal outcome variables** are also presented. We conclude with a discussion of model selection, which applies to models with two or more predictor variables.

For linear regression models with multiple predictors, in addition to our LINE assumptions, we add the assumption of no **multicollinearity**. That is, we assume our predictor variables are not themselves correlated.

`Rcmdr` and R have multiple ways to analyze linear regression models; we will continue to emphasize the **general linear model** approach, which allow us to handle continuous and categorical predictor variables.

Practical aspects of model diagnostics were presented in Chapter 17; these rules apply for multiple predictor variable models. Regression and correlation (Chapter 16) both test linear hypotheses: we state that the relationship between two variables is linear (the alternate hypothesis) or it is not (the null hypothesis). The difference? Correlation is a test of association (are variables correlated, we ask?), but are not tests of causation: we do not imply that one variable causes another to vary, even if the correlation between the two variables is large and positive, for example. Correlations are used in statistics on data sets not collected from explicit experimental designs incorporated to test specific hypotheses of cause and effect. Regression is to cause and effect as correlation is to association. Regression, ANOVA, and other general linear models are designed to permit the statistician to control for the effects of confounding variables provided the causal variables themselves are uncorrelated.

Models

Chapter 17 covered the simple linear model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Chapter 18 covers multiple regression linear model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon_i$$

where α or β_0 represent the Y-intercept and β or $\beta_1, \beta_2, \ldots \beta_n$ represent the regression slopes.

Chapter 18 also covers the logistic regression model

$$f(X) = \frac{L}{1 + e^{-k(X - X_0)}}$$

where L refers to the upper or maximum value of the curve, k refers to the rate of change at the steepest part of the curve, and X_0 refers to the inflection point of the curve. **Logistic functions** are S-shaped, and typical use involves looking at population growth rates (e.g., Fig. 18.1), or in the case of logistic regression, how a treatment affects the rate of growth.

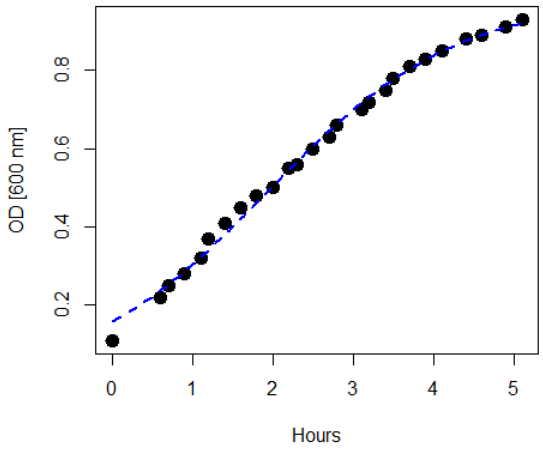


Figure 18.1: Growth of bacteria over time (optical density at 600 nm UV spectrophotometer), fit by logistic function (dashed line).

18.1: Multiple linear regression

Introduction

Last time we introduced simple linear regression:

- one independent X variable
- one dependent Y variable.

The linear relationship between Y and X was estimated by the method of **Ordinary Least Squares** (OLS). OLS minimizes the sum of squared distances between the observed responses, Y_i, and responses predicted by the line, $\hat{Y}_i$. Simple linear regression is analogous to our one-way ANOVA — one outcome or response variable and one factor or predictor variable (Chapter 12.2).

But the world is complicated and so, our one-way ANOVA was extended to the more general case of two or more predictor (factor) variables (Chapter 14). As you might have guessed by now, we can extend simple regression to include more than one predictor variable. In fact, combining ANOVA and regression gives you the **general linear model**! And, you should not be surprised that statistics has extended this logic to include not only multiple predictor variables, but also multiple response variables. Multiple response variables falls into a category of statistics called **multivariate statistics**.

Like multi-way ANOVA, multiple regression is the extension of simple linear regression from one independent predictor variable to include two or more predictors. The benefit of this extension is obvious — our models gain realism. All else being equal, the more predictors, the better the model will be at describing and/or predicting the response. Things are not all equal, of course, and we'll consider two complications of this basic premise, that more predictors are best; in some cases they are not.

However, before discussing the exceptions or even the complications of a multiple linear regression model, we begin by obtaining estimates of the full model, then introduce aspects of how to evaluate the model. We also introduce comparisons of models and whether a reduced model may be the preferred model.

R code

Multiple regression is easy to do in `Rcmdr` — recall that we used the general linear model function, `lm()`, to analyze **one-way ANOVA** and simple linear regression. In R Commander, we access `lm()` by

Rcmdr: Statistics → Fit model → Linear model

You may, however, access linear regression through R Commander

We use the same general linear model function for cases of multi-way ANOVA and for multiple regression problems. Simply enter more than one ratio-scale predictor variable and boom!

You now have yourself a multiple regression. You would then proceed to generate the ANOVA table for hypothesis testing

Rcmdr: Models → Hypothesis testing → ANOVA tables

From the output of the regression command, estimates of the coefficients along with standard errors for the estimate and results of t-tests for each coefficient against the respective null hypotheses for each coefficient are also provided. In our discussion of simple linear regression we introduced the components: the intercept, the slope, as well as the concept of model fit, as evidenced by R^2, the **coefficient of determination**. These components exist for the **multiple regression** problem, too, but now we call the slopes **partial regression slope**s because there are more than one.

Our full multiple regression model becomes

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon_i$$

where the coefficients $\beta_1, \beta_2, \ldots, \beta_n$ are the partial regression slopes and β_0 is the ***Y*-intercept** for a model with $1 - n$ predictor variables. Each coefficient has a null hypothesis, each has a standard error, and therefore, each coefficient can be tested by the t-test.

Now, regression, like ANOVA, is an enormous subject and we cannot do it justice in the few days we will devote to it. We can, however, walk you through a fairly typical example. I've posted a small data set `diabetesCholStatin` at the end of this page. Scroll down or click here. View the data set and complete your basic data exploration routine: make scatterplots and box plots. We think (predict) that body size and drug dose cause variation in serum cholesterol levels in adult men. But do both predict cholesterol levels?

Selecting the best model

We have two predictor variables, and we can start to see whether none, one, or both of the predictors contribute to differences in cholesterol levels. In this case, both contribute significantly. The power of multiple regression approaches is that it provides a simultaneous test of a model which may have many explanatory variables deemed appropriate to describe a particular response. More generally, it is sometimes advisable to think more philosophically about how to select a **best model**.

In model selection, some would invoke **Occam's razor** — given a set of explanations, the simplest should be selected — to justify seeking simpler models. There are a number of approaches (forward selection, backward selection, or stepwise selection), and the whole effort of deciding among competing models is complicated with a number of different assumptions, strengths and weaknesses. I refer you to the discussion below, which of course is just a very brief introduction to a very large subject in (bio)statistics!

Let's get the full regression model

The statistical model is

$$ChLDL_i = \beta_0 + \beta_1 \cdot BMI + \beta_2 \cdot Dose + \beta_3 \cdot Statin + \epsilon_i$$

As written in R format, our model is `ChLDL ~ BMI + Dose + Statin` .

> **Note:**
>
> `BMI` is ratio scale and `Statin` is categorical (two levels: Statin1, Statin2). `Dose` can be viewed as categorical, with five levels (5, 10, 20, 40, 80 mg), interval scale, or ratio scale. If we are make the assumption that the difference between 5, 10, up to 80 mg is meaningful, and that the effect of dose is at least proportional if not linear with respect to ChLDL, then we would treat `Dose` as ratio scale, not interval scale. That's what we did here.

We can now proceed in R Commander to fit the model.

Rmdr: Statistics → Fit models → Linear model

How the model is inputted into linear model menu is shown in Figure 18.1.1.

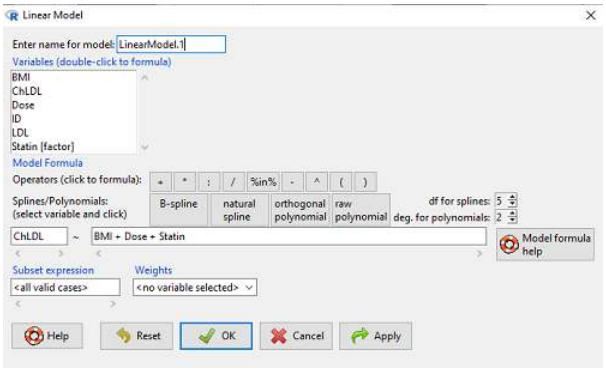


Figure 18.1.1: Screenshot of Rcmdr linear model menu with our model elements in place.

The output

```
summary(LinearModel.1)
Call:
lm(formula = ChLDL ~ BMI + Dose + Statin, data = cholStatins)

Residuals:
Min 1Q Median 3Q Max
-3.7756 -0.5147 -0.0449 0.5038 4.3821

Coefficients:
```

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.016715   1.178430   0.863 0.39041
BMI                   0.058078   0.047012   1.235 0.21970
Dose                 -0.014197   0.004829  -2.940 0.00411 **
Statin[Statin2]       0.514526   0.262127   1.963 0.05255 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.31 on 96 degrees of freedom
Multiple R-squared: 0.1231, Adjusted R-squared: 0.09565
F-statistic: 4.49 on 3 and 96 DF, p-value: 0.005407
```

Question. What are the estimates of the model coefficients (rounded)?

b_0 = intercept = 1.017

b_1 = slope for variable BMI = 0.058

b_2 = slope for variable Dose = -0.014

b_3 = slope for variable Statin = -0.515

Question. Which of the three coefficients were statistically different from their null hypothesis?

Answer: Only the b_2 coefficient was judged statistically significant at the Type I error level of 5% (p = 0.0041). Of the four null hypotheses we have for the coefficients (Intercept = 0; $b_1 = 0$; $b_2 = 0$; $b_3 = 0$), we only reject the null hypothesis for `Dose` coefficient.

Note the important concept about the lack of a direct relationship between the magnitude of the estimate of the coefficient and the likelihood that it will be statistically significant! In absolute value terms $b_1 > b_2$, but b_1 was not even close to statistical significance (p = 0.220).

We generate a 2D scatterplot and include the regression lines (by group=Statin) to convey the relationship between at least one of the predictors (Fig. 18.1.2).

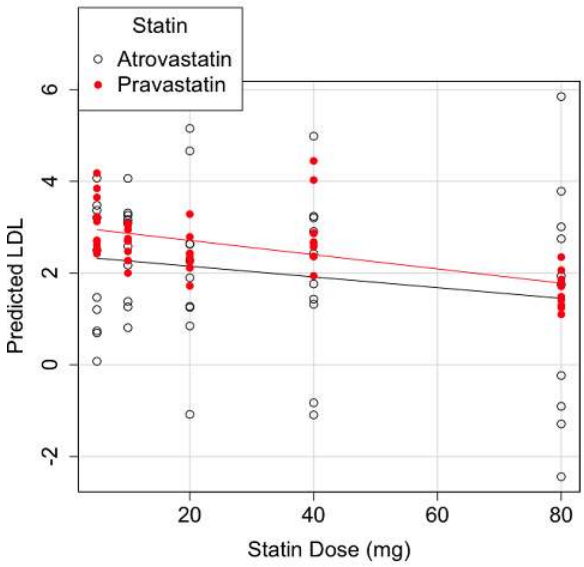


Figure \(\PageIndex{2}\): Scatter plot of predicted LDL against dose of a statin drug. Regression lines represent the different statin drugs (Statin1, Statin2).

Question. Based on the graph, can you explain why there will be no statistical differences between levels of the statin drug type, Statin1 (shown open circles) vs. Statin2 (shown closed red circles)?

Because we have two predictors (BMI and Statin Dose), you may also elect to use a 3D-scatterplot. Here's one possible result (Fig. 18.1.3).

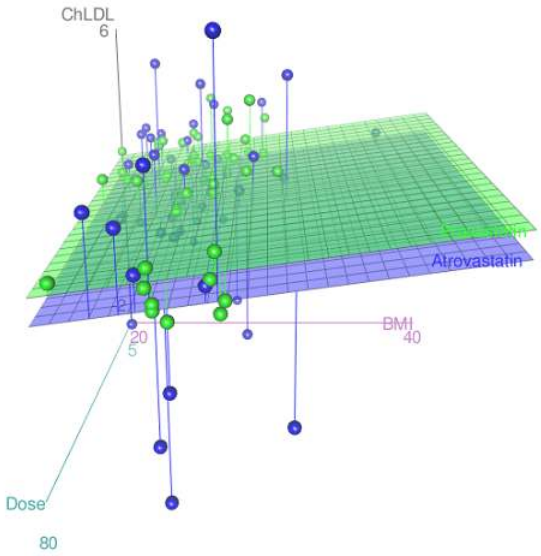


Figure 18.1.3: 3D plot of BMI and dose of Statin drugs on change in LDL levels (green Statin2, blue Statin1).

R code for Figure 18.1.3

Graph made in **Rcmdr: Graphs → 3D Graph → 3D scatterplot …**

```
scatter3d(ChLDL~BMI+Dose|Statin, data=diabetesCholStatin, fit="linear",
residuals=TRUE, parallel=FALSE, bg="white", axis.scales=TRUE, grid=TRUE,
ellipsoid=FALSE)
```

Note:

Figure 18.1.3 is a challenging graphic to interpret. I wouldn't use it because it doesn't convey a strong message. With some effort we can see the two planes representing mean differences between the two statin drugs across all predictors, but it's a stretch. No doubt the graph can be improved by changing colors, for example, but I think the 2d plot (Figure 18.1.2) works better. Alternatively, if the platform allows, you can use animation options to help your reader see the graph elements. Interactive graphics are very promising and, again, unsurprisingly, there are several R packages available. For this example, `plot3d()` of the package `rgl` can be used. Figure 18.1.4 is one possible version; I saved images and made animated gif.

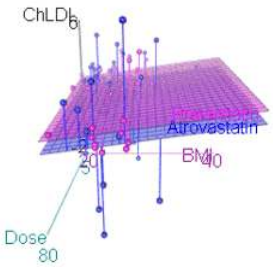


Figure 18.1.4: An example of a possible interactive 3D plot; the file embedded in this page is not interactive, just an animation.

Diagnostic plots

While visualization concerns are important, let's return to the statistics. All evaluations of regression equations should involve an inspection of the residuals. Inspection of the residuals allows you to decide if the regression fits the data; if the fit is adequate, you

then proceed to evaluate the statistical significance of the coefficients.

The default diagnostic plots (Fig. 18.1.5) R provides are available from **Rcmdr: Models → Graphs → Basic diagnostics plots**

Four plots are returned:

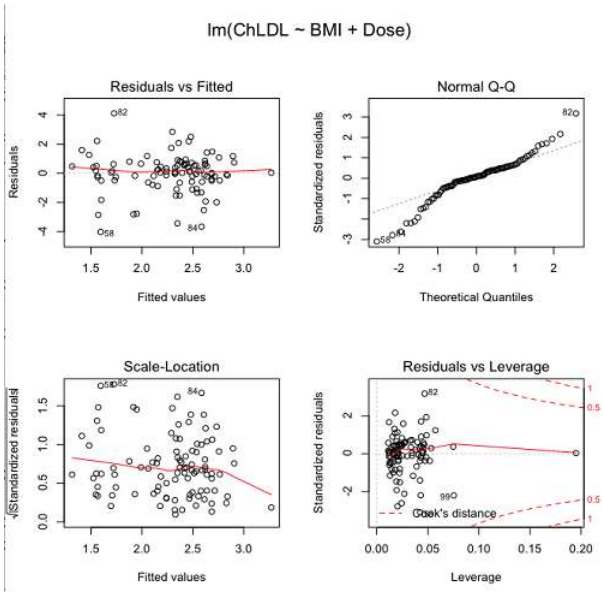


Figure 18.1.5: R's default regression diagnostic plots.

Each of these diagnostic plots in Figure 18.1.5 gives you clues about the model fit.

1. Plot of **residuals vs. fitted** helps you identify patterns in the residuals
2. **Normal Q-Q** plot helps you to see if the residuals are approximately normally distributed
3. **Scale-location** plot provides a view of the spread of the residuals
4. The **residuals vs. leverage** plot allows you to identify influential data points.

We introduced these plots in Chapter 17.8 when we discussed fit of simple linear model to data. My conclusion? No obvious trend in residuals, so linear regression is a fit to the data; data not normally distributed, as Q-Q plot shows S-shape.

Interpreting the diagnostic plots for this problem

The "**Normal Q-Q**" plot allows us to view our residuals against a normal distribution (the dotted line). Our residuals do no show an ideal distribution: low for the first quartile, about on the line for intermediate values, then high for the 3rd and 4th quartile residuals. If the data were bivariate normal we would see the data fall along a straight line. The "S-shape" suggests log-transformation of the response and or one or more of the predictor variables.

Note that there also seems to be a pattern in **residuals vs the predicted** (fitted) values. There is a trend of increasing residuals as cholesterol levels increase, which is particularly evident in the "**scale-location**" plot. Residuals tended to be positive at low and high doses, but negative at intermediate doses. This suggests that the relationship between predictors and cholesterol levels may not be linear, and it demonstrates what statisticians refer to as a monotonic spread of residuals.

The last diagnostic plot looks for individual points that influence, change, or "**leverage**" the regression — in other words, if a point is removed, does the general pattern change? If so, then the point had "leverage" and thus we need to decide whether or not to include the datum. diagnostic plots **Cook's distance** is a measure of the influence of a point in regression. Points with large Cook's distance values warrant additional checking.

The multicollinearity problem

Statistical model building is a balancing act by the statistician. While simpler models may be easier to interpret and, perhaps, to use, it is a basic truism that the more predictor variables the model includes, the more realistic the statistical model. However, each additional parameter that is added to the statistical model must be independent of all other parameters already in the model. To the extent that this assumption is violated, the problem is termed **multicollinearity**. If predictor variables are highly correlated, then they are essentially just linear combinations and do not provide independent evidence. For example, one would naturally not

include two core body temperature variables in a statistical model on basal metabolic rate, one in degrees Fahrenheit and the other in degrees Celsius, because it is a simple linear conversion between the two units. This would be an example of **structural collinearity**: the collinearity is because of misspecification of the model variables. In contrast, collinearity among predictor variables may because the data are themselves correlated. For example, if multiple measures of body size are included (weight, height, length of arm, etc.), then we would expect these to be correlated, i.e., **data multicollinearity**.

Collinearity in statistical models may have a number of undesirable effects on a multiple regression model. These include

- estimates of coefficients not stable: with collinearity, values of coefficients depend on other variables in the model; if collinear predictors, then the assumption of independent predictor variables is violated.
- precision of the estimates decreases (standard error of estimates increase).
- statistical power decreases.
- p-values for individual coefficients not trust worthy.

Tolerance and Variance Inflation Factor

Absence of multicollinearity is important assumption of multiple regression. A partial test is to calculate product moment correlations among predictor variables. For example, when we calculate the correlation between BMI and Dose for our model, we get $r = 0.101$ $(p = 0.3186)$, and therefore would tentatively conclude that there was little correlation between our predictor variables.

A number of diagnostic statistics have been developed to test for multicollinearity. **Tolerance** for a particular independent variable (X_i) is defined as 1 minus the proportion of variance it shares with the other independent variables in the regression analysis $\left(1 - R_i^2\right)$ (O'Brien 2007). Tolerance reports the proportion of total variance explained by adding the X_i^{th} predictor variable that is unrelated to the other variables in the model. A small value for tolerance indicates multicollinearity — and that the predictor variable is nearly a perfect combination (linear) of the variables already in the model and therefore should be omitted from the model. Because tolerance is defined in relation to the coefficient of determination, you can interpret a tolerance score as the unique variance accounted for by a predictor variable.

A second, related diagnostic of multicollinearity is called the Variance Inflation Factor, **VIF**. VIF is the inverse of tolerance.

$$VIF = \frac{1}{tolerance}$$

VIF shows how much of the variance of a regression coefficient is increased because of collinearity with the other predictor variables in the model. VIF is easy to interpret: a tolerance of 0.01 has a VIF of 100; a tolerance of 0.1 has a VIF of 10; a tolerance of 0.5 has a VIF of 2, and so on. Thus, small values of tolerance and large values of VIF are taken as evidence of multicollinearity.

Rcmdr: Models → Numerical diagnostics → Variation-inflation factors

```
vif(RegModel.2)
      BMI     Dose
1.010256 1.010256
```

A rule of thumb is that if VIF is greater than 5 then there is multicollinearity; with VIF values close to one we would conclude, like our results from the partial correlation estimate above, that there is little evidence for a problem of collinearity between the two predictor variables. They can therefore remain in the model.

Solutions for multicollinearity

If there is substantial multicollinearity then you cannot simply trust the estimates of the coefficients. Assuming that there hasn't been some kind of coding error on your part, then you may need to find a solution. One solution is to drop one of the predictor variables and redo the regression model. Another option is to run what is called a Principle Components Regression. One takes the predictor variables and runs a Principle Component Analysis to reduce the number of variables, then the regression is run on the PCA components. By definition, the PCA components are independent of each other. Another option is to use ridge regression approach.

Like any diagnostic rule, however, one should not blindly apply a rule of thumb. A VIF of 10 or more may indicate multicollinearity, but it does not necessarily lead to the conclusion that the linear regression model requires that the researcher

reduce the number of predictor variables or analyze the problem using a different statistical method to address multicollinearity as the sole criteria of a poor statistical model. Rather, the researcher needs to address all of the other issues about model and parameter estimate stability, including sample size. Unless the collinearity is extreme (like a correlation of 1.0 between predictor variables!), larger sample sizes alone will work in favor of better model stability (by lowering the sample error) (O'Brien 2007).

Questions

1. Can you explain why the magnitude of the slope is not the key to statistical significance of a slope? Hint: look at the equation of the t-test for statistical significance of the slope.
2. Consider the following scenario. A researcher repeatedly measures his subjects for blood pressure over several weeks, then plots all of the values over time. In all, the data set consists of thousands of readings. He then proceeds to develop a model to explain blood pressure changes over time. What kind of collinearity is present in his data set? Explain your choice.
3. We noted that `Dose` could be viewed as categorical variable. Convert `Dose` to factor variable (`fDose`) and redo the linear model. Compare the summary output and discuss the additional coefficients.
   - Use **Rcmdr: Data → Manage variables in active data set → Convert numeric Variables to Factors** to create a new factor variable `fDose` . It's ok to use the numbers as factor levels.
4. We flagged the change in LDL as likely to be not normally distributed. Create a $\log_{10}$-transformed variable for `ChLDL` and perform the multiple regression again.
   a. Write the new statistical model
   b. Obtain the regression coefficients — are they statistically significant?
   c. Run basic diagnostic plots and evaluate for fit of the linear model for this data set.

Data set

ID	Statin	Dose	BMI	LDL	ChLDL
1	Statin2	5	19.5	3.497	2.7147779309
2	Statin1	20	20.2	4.268	1.2764831106
3	Statin2	40	20.3	3.989	2.6773769532
4	Statin2	20	20.3	3.502	2.4306181501
5	Statin2	80	20.4	3.766	1.7946303961
6	Statin2	20	20.6	3.44	2.2342950639
7	Statin1	20	20.7	3.414	2.6353051933
8	Statin1	10	20.8	3.222	0.8091810801
9	Statin1	10	21.1	4.04	3.2595985907
10	Statin1	40	21.2	4.429	1.7639974729
11	Statin1	5	21.2	3.528	3.3693768458
12	Statin1	40	21.5	3.01	-0.8271542022
13	Statin2	20	21.6	3.393	2.1117204833
14	Statin1	10	21.7	4.512	3.1662377996
15	Statin1	80	22	5.449	3.0083296182
16	Statin2	10	22.2	4.03	3.0501301624
17	Statin2	40	22.2	3.911	2.6460344888
18	Statin2	10	22.2	3.724	2.9456555243
19	Statin1	5	22.2	3.238	3.2095842825

20	Statin2	10	22.5	4.123	3.0887629267
21	Statin1	20	22.6	3.859	5.1525478688
22	Statin1	10	23	4.926	2.58482964
23	Statin2	20	23	3.512	2.2919748394
24	Statin1	5	23	3.838	1.4689995606
25	Statin2	20	23.1	3.548	2.3407899756
26	Statin1	5	23.1	3.424	1.2043457967
27	Statin1	40	23.2	3.709	3.2381790892
28	Statin1	80	23.2	4.786	2.7486432463
29	Statin1	20	23.3	4.103	1.2500819426
30	Statin1	40	23.4	3.341	1.4322916002
31	Statin1	10	23.5	3.828	1.3817551192
32	Statin2	10	23.8	4.02	3.0391874265
33	Statin1	20	23.8	3.942	0.8483284736
34	Statin2	20	23.8	2.89	1.7211634664
35	Statin1	80	23.9	3.326	1.9393460444
36	Statin1	10	24.1	4.071	3.0907410326
37	Statin1	40	24.1	4.222	1.3223045884
38	Statin2	10	24.1	3.44	2.472222941
39	Statin1	5	24.2	3.507	0.0768171794
40	Statin2	20	24.2	3.647	2.4257575585
41	Statin2	80	24.3	3.812	1.7105748759
42	Statin2	40	24.3	3.305	1.9405724055
43	Statin2	5	24.3	3.455	2.5022137646
44	Statin2	5	24.4	4.258	3.2280077893
45	Statin1	5	24.4	4.16	3.4777470262
46	Statin2	80	24.4	4.128	2.0632471844
47	Statin1	80	24.5	4.507	3.784421647
48	Statin1	5	24.5	3.553	0.6957091748
49	Statin2	10	24.5	3.616	2.6998703189
50	Statin2	80	24.6	3.372	1.3004010967
51	Statin2	80	24.6	3.667	1.4181086606
52	Statin2	5	24.7	3.854	3.1266706892
53	Statin1	80	24.7	3.32	-1.2864388279
54	Statin2	5	24.7	3.756	2.4236635094

55	Statin1	40	24.8	4.398	2.907472945
56	Statin2	40	24.9	3.621	2.3624285593
57	Statin1	10	25	3.17	1.264656476
58	Statin1	80	25.1	3.424	-2.4369077381
59	Statin2	10	25.1	3.196	2.0014648648
60	Statin2	80	25.2	3.367	1.1007041451
61	Statin1	80	25.2	3.067	-0.2315398019
62	Statin1	20	25.3	3.678	4.6628661348
63	Statin2	5	25.5	4.077	2.6117051224
64	Statin1	20	25.5	3.678	2.6330531096
65	Statin2	5	25.6	4.994	4.1800816149
66	Statin1	20	25.8	3.699	1.8990314684
67	Statin1	10	25.9	3.507	4.0637570533
68	Statin2	20	25.9	3.445	2.3037613081
69	Statin1	5	26	4.025	2.50142676
70	Statin1	5	26.3	3.616	0.7408631019
71	Statin2	40	26.4	3.937	2.5733214297
72	Statin2	40	26.4	3.823	2.3638394785
73	Statin1	10	26.7	4.46	2.1741977546
74	Statin2	5	26.7	5.03	3.845271327
75	Statin2	10	26.7	3.73	2.7088955103
76	Statin2	10	26.7	3.232	2.2726268196
77	Statin1	80	26.8	3.693	1.751169214
78	Statin2	80	27	4.108	1.8613104992
79	Statin2	40	27.2	5.398	4.0289773539
80	Statin2	80	27.2	4.517	2.3489030399
81	Statin2	20	27.3	3.901	2.7900467077
82	Statin1	80	27.3	5.247	5.8485450123
83	Statin2	80	27.4	3.507	1.2478629747
84	Statin1	20	27.4	3.807	-1.0799279924
85	Statin2	80	27.6	3.574	1.48678931
86	Statin1	40	27.8	4.16	2.4277532799
87	Statin2	20	28	4.501	3.2846482963
88	Statin2	5	28.1	3.621	2.6990067113
89	Statin1	40	28.2	3.652	-1.0912561688

90	Statin2	40	28.2	4.191	2.8742307203
91	Statin2	40	28.4	5.791	4.4454535731
92	Statin1	40	28.6	4.698	3.2028737773
93	Statin1	5	29	4.32	4.0707532197
94	Statin2	10	29.1	3.776	2.7512805004
95	Statin2	5	29.2	4.703	3.6494895215
96	Statin2	40	29.9	4.128	2.8646910266
97	Statin1	40	30.4	4.693	4.9837039826
98	Statin1	20	30.4	4.123	2.2738979752
99	Statin1	80	30.5	3.921	-0.9034376511
100	Statin1	10	36.5	4.175	3.3114366758

18.2: Nonlinear regression

Introduction

The **linear model** is incredibly relevant in so many cases. A quick look for "linear model" in PUBMED returns about 22 thousand hits; 3.7 million in Google Scholar; 3 thousand hits in ERIC database. These results compare to search of "statistics" in the same databases: 2.7 million (PUBMED), 7.8 million (Google Scholar), 61.4 thousand (ERIC). But all models are not the same.

Fit of a model to the data can be evaluated by looking at the plots of residuals (Fig. 18.2.1), where we expect to find **random distribution** of **residuals** across the range of predictor variable.

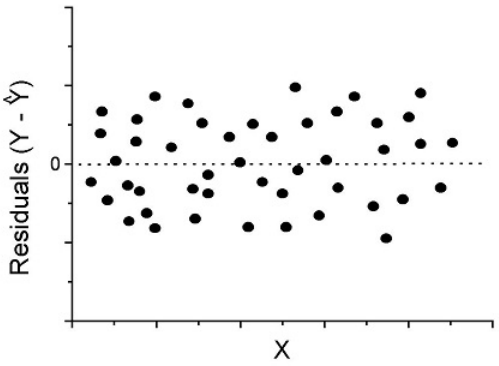


Figure 18.2.1: Ideal plot of residuals against values of X, the predictor variable, for a well-supported linear model fit to the data.

However, clearly, there are problems for which assumption of fit to line is not appropriate. We see this, again, in **patterns of residuals**, e.g., Figure 18.2.2

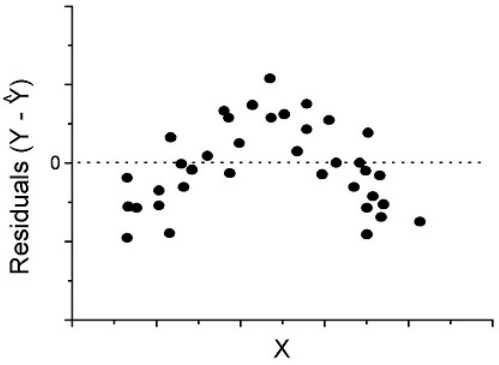


Figure 18.2.2: Example of residual plot; pattern suggests nonlinear fit.

Fitting of polynomial linear model

Fit simple linear regression, using data linked at end of page. Data sourced from Yuan et al. (2012), https://phenome.jax.org/projects/Yuan2.

R code:

```
LinearModel.1 <- lm(cumFreq~Months, data=yuan)

summary(LinearModel.1)

Call:
```

```
lm(formula = cumFreq ~ Months, data = yuan)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11070 -0.07799 -0.01728 0.06982 0.13345

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.132709   0.045757   -2.90  0.0124 *
Months           0.029605   0.001854   15.97  6.37e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09308 on 13 degrees of freedom
Multiple R-squared: 0.9515, Adjusted R-squared: 0.9477
F-statistic: 254.9 on 1 and 13 DF, p-value: 6.374e-10
```

We see from the R^2 (95%), a high degree of fit to the data. However, residual plot reveals obvious trend (Fig. 18.2.3)

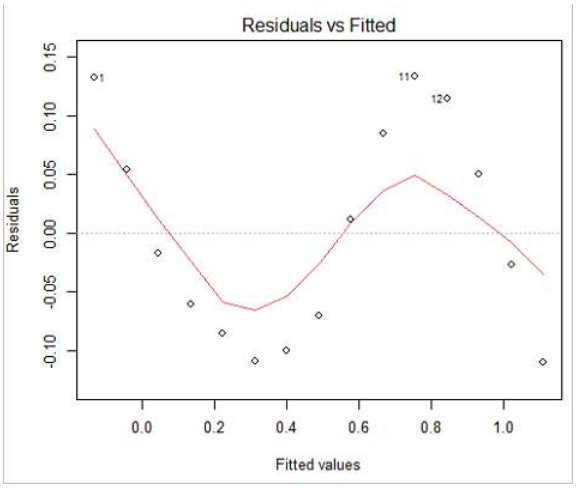


Figure 18.2.3: Residual plot.

We can fit a polynomial regression.

First, a second order polynomial:

```
LinearModel.2 <- lm(cumFreq ~ poly( Months, degree=2), data=yuan)

summary(LinearModel.2)

Call:
lm(formula = cumFreq ~ poly(Months, degree = 2), data = yuan)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13996 -0.06720 -0.02338 0.07153 0.14277

Coefficients:
```

```
                            Estimate Std. Error  t value   Pr(>|t|)
(Intercept)                  0.48900    0.02458   19.891   1.49e-10 ***
poly(Months, degree = 2)1    1.48616    0.09521   15.609   2.46e-09 ***
poly(Months, degree = 2)2    0.06195    0.09521    0.651   0.528
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09521 on 12 degrees of freedom
Multiple R-squared: 0.9531, Adjusted R-squared: 0.9453
F-statistic: 122 on 2 and 12 DF, p-value: 0.0000000106
```

Second, try a third order polynomial:

```
LinearModel.3 <- lm(cumFreq ~ poly(Months, degree = 3), data=yuan)

summary(LinearModel.3)

Call:
lm(formula = cumFreq ~ poly(Months, degree = 3), data = yuan)

Residuals:
      Min        1Q    Median       3Q      Max
-0.052595 -0.021533 0.001023 0.025166 0.048270

Coefficients:
                            Estimate Std. Error t value   Pr(>|t|)
(Intercept)                 0.488995   0.008982  54.442   9.90e-15 ***
poly(Months, degree = 3)1   1.486157   0.034787  42.722   1.41e-13 ***
poly(Months, degree = 3)2   0.061955   0.034787   1.781   0.103
poly(Months, degree = 3)3 -0.308996   0.034787  -8.883   2.38e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03479 on 11 degrees of freedom
Multiple R-squared: 0.9943, Adjusted R-squared: 0.9927
F-statistic: 635.7 on 3 and 11 DF, p-value: 1.322e-12
```

Which model is best? We are tempted to compare R-squared among the models, but R^2 turn out to be untrustworthy here. Instead, we compare using the **Akaike Information Criterion, AIC**

R code/results:

```
AIC(LinearModel.1,LinearModel.2, LinearModel.3)
              df       AIC
LinearModel.1  3 -24.80759
LinearModel.2  4 -23.32771
LinearModel.3  5 -52.83981
```

Smaller the AIC, better fit.

```
anova(RegModel.5,LinearModel.3, LinearModel.4)
Analysis of Variance Table

Model 1: cumFreq ~ Months
Model 2: cumFreq ~ poly(Months, degree = 2)
Model 3: cumFreq ~ poly(Months, degree = 3)
  Res.Df   RSS    Df Sum of Sq       F      Pr(>F)
1     13 0.112628
2     12 0.108789  1  0.003838  3.1719 0.1025
3     11 0.013311  1  0.095478 78.9004 0.000002383 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic regression

The Logistic regression is a classic example of nonlinear model.

R code

```
logisticModel <-nls(cumFreq~DD/(1+exp(-(CC+bb*Months))), start=list(DD=1,CC=0.2,bb=.5
5.163059 : 1.0 0.2 0.5
2.293604 : 0.90564552 -0.07274945 0.11721201
1.109135 : 0.96341283 -0.60471162 0.05066694
0.429202 : 1.29060000 -2.09743525 0.06785993
0.3863037 : 1.10392723 -2.14457296 0.08133307
0.2848133 : 0.9785669 -2.4341333 0.1058674
0.1080423 : 0.9646295 -3.1918526 0.1462331
0.005888491 : 1.0297915 -4.3908114 0.1982491
0.004374918 : 1.0386521 -4.6096564 0.2062024
0.004370212 : 1.0384803 -4.6264657 0.2068853
0.004370201 : 1.0385065 -4.6269276 0.2068962
0.004370201 : 1.0385041 -4.6269822 0.2068989
```

```
summary(logisticModel)

Formula: yuan$cumFreq ~ DD/(1 + exp(-(CC + bb * yuan$Months)))

Parameters:
      Estimate  Std. Error  t value      Pr(>|t|)
DD    1.038504    0.014471    71.77   < 2e-16 ***
CC   -4.626982    0.175109   -26.42  5.29e-12 ***
bb    0.206899    0.008777    23.57  2.03e-11 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01908 on 12 degrees of freedom
```

```
Number of iterations to convergence: 11
Achieved convergence tolerance: 0.000006909

> AIC(logisticModel)
[1] -71.5467
```

Logistic regression is a statistical method for modeling the dependence of a **categorical (binomial) outcome variable** on one or more categorical and continuous predictor variables (Bewick et al 2005).

The **logistic function** is used to transform a sigmoidal curve to a more or less straight line while also changing the range of the data from binary (0 to 1) to infinity $(-\infty, +\infty)$. For event with probability of occurring p, the logistic function is written as

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

where ln refers to the **natural logarithm**.

This is an **odds ratio**. It represents the effect of the predictor variable on the chance that the event will occur.

The logistic regression model then very much resembles the same as we have seen before.

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

In R and `Rcmdr` we use the `glm()` function to model the logistic function. Logistic regression is used to model a binary outcome variable. What is a binary outcome variable? It is categorical! Examples include: Living or Dead; Diabetes Yes or No; Coronary artery disease Yes or No. Male or Female. One of the categories could be scored 0, the other scored 1. For example, living might be 0 and dead might be scored as 1. (By the way, for a binomial variable, the mean for the variable is simply the number of experimental units with "1" divided by the total sample size.)

With the addition of a binary response variable, we are now really close to the **Generalized Linear Model**. Now we can handle statistical models in which our predictor variables are either categorical or ratio scale. All of the rules of crossed, balanced, nested, blocked designs still apply because our model is still of a linear form.

We write our generalized linear model

$$G \sim Model$$

just to distinguish it from a general linear model with the ratio-scale Y as the response variable.

Think of the logistic regression as modeling a **threshold of change** between the 0 and the 1 value. In another way, think of all of the processes in nature in which there is a slow increase, followed by a rapid increase once a transition point is met, only to see the rate of change slow down again. Growth is like that. We start small, stay relatively small until birth, then as we reach our early teen years, a rapid change in growth (height, weight) is typically seen (well, not in my case … at least for the height). The fitted curve I described is a logistic one (other models exist too). Where the linear regression function was used to minimize the squared residuals as the definition of the best fitting line, now we use the logistic as one possible way to describe or best fit this type of a curved relationship between an outcome and one or more predictor variables. We then set out to describe a model which captures when an event is unlikely to occur (the probability of dying is close to zero) AND to also describe when the event is highly likely to occur (the probability is close to one).

A simple way to view this is to think of time being the predictor (X) variable and risk of dying. If we're talking about the lifetime of a mouse (lifespan typically about 18-36 months), then the risk of dying at one month is very low, and remains low through adulthood until the mouse begins the aging process. Here's what the plot might look like, with the probability of dying at age X on the Y axis (probability = 0 to 1) (Fig. 18.2.4).

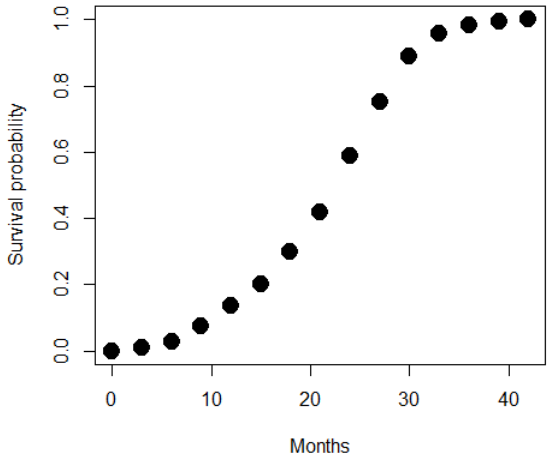


Figure 18.2.4: Lifespan of 1881 mice from 31 inbred strains (Data from Yuan et al (2012) available at https://phenome.jax.org/projects/Yuan2).

We ask — of all the possible models we could draw, which best fits the data? The curve fitting process is called the logistic regression.

With some minor, but important differences, running the logistic regression is the same as what you have been doing so far for ANOVA and for linear regression. In Rcmdr, access the logistic regression function by invoking the Generalized Linear Model (Fig. 5).

Rcmdr: Statistics → Fit models → Generalized linear model.

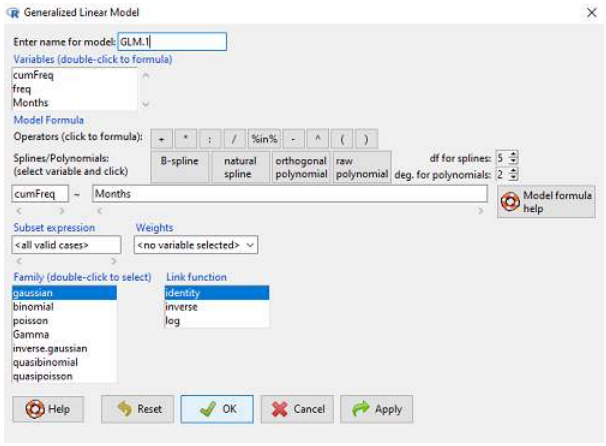


Figure 18.2.5: Screenshot of Rcmdr GLM menu. For logistic on ration-scale dependent variable, select gaussian family and identity link function.

Select the model as before. The box to the left accepts your binomial dependent variable; the box at right accepts your factors, your interactions, and your covariates. It permits you to inform R how to handle the factors: Crossed? Just enter the factors and follow each with a plus. If fully crossed, then the interactions may be specified with ":" to explicitly call for a two-way interaction between two (A:B) or a three-way interaction between three (A:B:C) variables. In the later case, if all of the two way interactions are of interest, simply typing A*B*C would have done it. If nested, then use %in% to specify the nesting factor.

R output:

```
> GLM.1 <- glm(cumFreq ~ Months, family=gaussian(identity), data=yuan)

> summary(GLM.1)

Call:
glm(formula = cumFreq ~ Months, family = gaussian(identity),
```

```
data = yuan)

Deviance Residuals:
Min 1Q Median 3Q Max
-0.11070 -0.07799 -0.01728 0.06982 0.13345

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.132709 0.045757 -2.90 0.0124 *
Months 0.029605 0.001854 15.97 6.37e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.008663679)

Null deviance: 2.32129 on 14 degrees of freedom
Residual deviance: 0.11263 on 13 degrees of freedom
AIC: -24.808

Number of Fisher Scoring iterations: 2
```

Assessing fit of the logistic regression model

Some of the differences you will see with the logistic regression is the term **deviance**. Deviance in statistics simply means compare one model to another and calculate some test statistic we'll call "the deviance." We then evaluate the size of the deviance like a chi-square goodness of fit. If the model fits the data poorly (residuals large relative to the predicted curve), then the deviance will be small and the probability will also be high — the model explains little of the data variation. On the other hand, if the deviance is large, then the probability will be small — the model explains the data, and the probability associated with the deviance will be small (significantly so? You guessed it! $P < 0.05$).

The **Wald test statistic** is

$$\left(\frac{\beta_n}{SE_{\beta_n}}\right)^2$$

where n and β refer to any of the n coefficients from the logistic regression equation and SE refers to the standard error if the coefficient. The Wald test is used to test the statistical significance of the coefficients. It is distributed approximately as a chi-squared probability distribution with one degree of freedom. The Wald test is reasonable, but has been found to give values that are not possible for the parameter (e.g., negative probability).

Likelihood ratio tests are generally preferred over the Wald test. For a coefficient, the likelihood test is written as

$$-2 \times \ln(likelihood\ ratio) = -2\ln(L_0/L_1) = -2 \times (\ln L_0 - \ln L_1)$$

where L_0 is the likelihood of the data when the coefficient is removed from the model (i.e., set to zero value), whereas L_1 is the likelihood of the data when the coefficient is the estimated value of the coefficient. It is also distributed approximately as a chi-squared probability distribution with one degree of freedom.

Questions

[pending]

Data set

Months	freq	cumFreq

0	0	0
3	0.01063264221159	0.01063264221159
6	0.017012227538543	0.027644869750133
9	0.045188729399256	0.072833599149389
12	0.064327485380117	0.137161084529506
15	0.064859117490697	0.202020202020202
18	0.097820308346624	0.299840510366826
21	0.118553960659224	0.41839447102605
24	0.171185539606592	0.589580010632642
27	0.162147793726741	0.751727804359383
30	0.137161084529506	0.888888888888889
33	0.069643806485912	0.958532695374801
36	0.024455077086656	0.982987772461457
39	0.011695906432749	0.994683678894205
42	0.005316321105795	1

18.3: Logistic regression

Introduction

We briefly introduced logistic regression in the previous chapter on nonlinear regression. We expand our discussion of logistic regression here.

Logistic regression is a statistical method for modeling the dependence of a categorical (binomial) outcome variable on one or more categorical and continuous predictor variables (Bewick et al 2005).

The logistic function may be used to transform a sigmoidal curve to a more or less straight line while also changing the range of the data from binary (0 to 1) to infinity $(-\infty, +\infty)$. For event with probability of occurring p, the logistic function is written as

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

where ln refers to the natural logarithm.

This is an odds ratio. It represents the effect of the predictor variable on the chance that the event will occur.

The logistic regression model then very much resembles the same general linear models we have seen before.

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

In R and `Rcmdr` we use the `glm()` function to model the logistic function. Logistic regression is used to model a binary outcome variable. What is a binary outcome variable? It is categorical! Examples include: Living or Dead; Diabetes Yes or No; Coronary artery disease Yes or No. Male or Female. One of the categories could be scored 0, the other scored 1. For example, living might be 0 and dead might be scored as 1. (By the way, for a binomial variable, the mean for the variable is simply the number of experimental units with "1" divided by the total sample size.)

With the addition of a binary response variable, we are now really close to the Generalized Linear Model. Now we can handle statistical models in which our predictor variables are either categorical or ratio scale. All of the rules of crossed, balanced, nested, blocked designs still apply because our model is still of a linear form.

We write our generalized linear model

$$G \sim Model$$

just to distinguish it from a general linear model with the ratio-scale Y as the response variable.

Think of the logistic regression as modeling a threshold of change between the 0 and the 1 value. In another way, think of all of the processes in nature in which there is a slow increase, followed by a rapid increase once a transition point is met, only to see the rate of change slow down again. Growth is like that (see Chapter 20.10 for related growth and related models). We start small, stay relatively small until birth, then as we reach our early teen years, a rapid change in growth (height, weight) is typically seed (well, not in my case … at least for the height). The curve I described is a logistic one (other models exist too). Where the linear regression function was used to minimize the squared residuals as the definition of the best fitting line, now we use the logistic as one possible way to describe or best fit this type of a curved relationship between an outcome and one or more predictor variables. We then set out to describe a model which captures when an event is unlikely to occur (the probability of dying is close to zero) AND to also describe when the event is highly likely to occur (the probability is close to one).

A simple way to view this is to think of time being the predictor (X) variable and risk of dying. If we're talking about the lifetime of a mouse (lifespan typically about 18-36 months), then the risk of dying at one months is very low, and remains low through adulthood until the mouse begins the aging process. Here's what the plot might look like, with the probability of dying at age X on the Y axis (probability = 0 to 1) (Fig. 18.3.1).

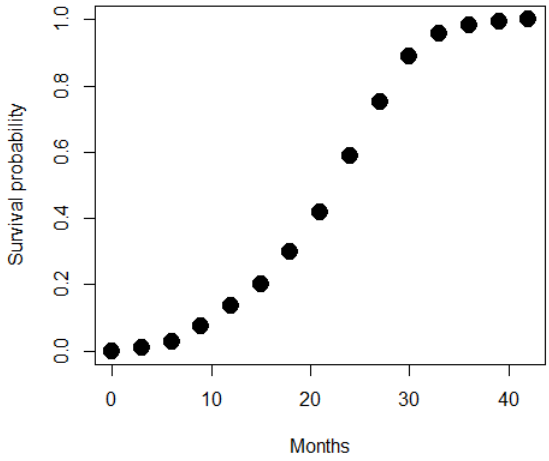


Figure 18.3.1: Lifespan of 1881 mice from 31 inbred strains (Data from Yuan et al [2012] available at https://phenome.jax.org/projects/Yuan2). *Note: I labeled Y axis labeled "Survival Probability"; "Inverse Survival Probability" would be more accurate.*

We ask — of all the possible models we could draw — which model best fits the data? The curve fitting process is called the logistic regression. The sample data set is listed at end of this page (scroll down or click here). Create **data.frame** called yuan.

With some minor, but important differences, running the logistic regression is the same as what you have been doing so far for ANOVA and for linear regression. In Rcmdr, access the **logistic regression function** by calling the **Generalized Linear Model** (Fig. 18.3.2).

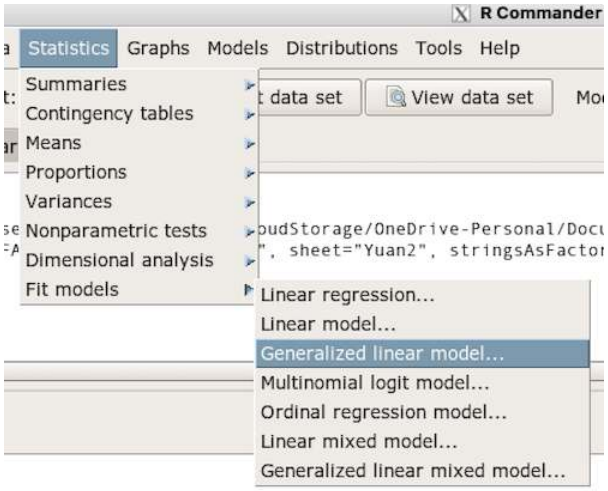


Figure 18.3.2: Access Generalized Linear Model via R Commander.

R results:

```
GLM.1 <- glm(cumFreq ~ Months, family=gaussian(identity), data=yuan)

> summary(GLM.1)

Call:
glm(formula = cumFreq ~ Months, family = gaussian(identity),
data = yuan)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.11070 -0.07799 -0.01728 0.06982 0.13345
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.132709   0.045757   -2.90 0.0124 *
Months       0.029605   0.001854   15.97 6.37e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.008663679)

Null deviance: 2.32129 on 14 degrees of freedom
Residual deviance: 0.11263 on 13 degrees of freedom
AIC: -24.808

Number of Fisher Scoring iterations: 2
```

Rcmdr: Statistics → Fit models → Generalized linear model.

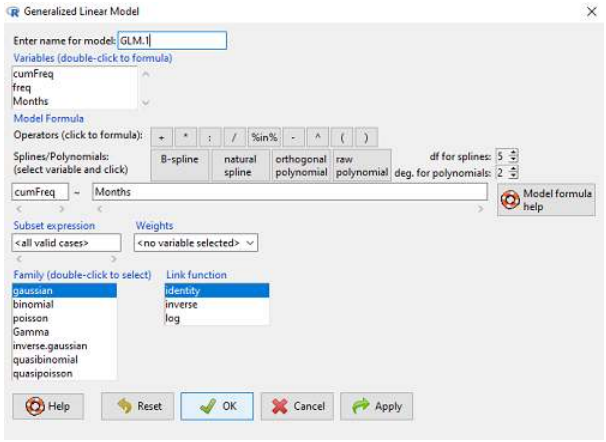


Figure 18.3.3: Screenshot of Rcmdr GLM menu. For logistic on ratio-scale dependent variable, select gaussian family and identity link function.

Select the model as before. The box to the left accepts your binomial dependent variable; the box at right accepts your factors, your interactions, and your covariates. It permits you to inform R how to handle the factors: Crossed? Just enter the factors and follow each with a plus. If fully crossed, then the interactions may be specified with ":" to explicitly call for a two-way interaction between two (A:B) or a three-way interaction between three (A:B:C) variables. In the later case, if all of the two way interactions are of interest, simply typing A*B*C would have done it. If nested, then use %in% to specify the nesting factor.

R output:

```
GLM.1 <- glm(cumFreq ~ Months, family=gaussian(identity), data=yuan)

summary(GLM.1)

Call:
glm(formula = cumFreq ~ Months, family = gaussian(identity),
data = yuan)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.11070 -0.07799 -0.01728 0.06982 0.13345

Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   -0.132709     0.045757   -2.90     0.0124 *
Months         0.029605     0.001854   15.97     6.37e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.008663679)

Null deviance: 2.32129 on 14 degrees of freedom
Residual deviance: 0.11263 on 13 degrees of freedom
AIC: -24.808

Number of Fisher Scoring iterations: 2
```

Assessing fit of the logistic regression model

Some of the differences you will see with the logistic regression is the term "deviance." deviance in statistics simply means compare one model to another and calculate some test statistic we'll call "the deviance." We then evaluate the size of the deviance like a chi-square goodness of fit. If the model fits the data poorly (residuals large relative to the predicted curve), then the deviance will be small and the probability will also be high — the model explains little of the data variation. On the other hand, if the deviance is large, then the probability will be small — the model explains the data, and the probability associated with the deviance will be small (significantly so? You guessed it! P < 0.05).

The Wald statistic is

$$\left(\frac{\beta_n}{SE_{\beta_n}}\right)^2$$

where n and β refer to any of the n coefficient from the logistic regression equation and SE refers to the standard error if the coefficient. The Wald test is used to test the statistical significance of the coefficients. It is distributed approximately as a chi-squared probability distribution with one degree of freedom. The Wald test is reasonable, but has been found to give values that are not possible for the parameter (e.g., negative probability).

Likelihood ratio tests are generally preferred over the Wald test. For a coefficient, the likelihood test is written as

$$-2 \times \ln(likelihood\ ratio) = -2\ \ln(L_0/L_1) = -2 \times (\ln L_0 - \ln L_1)$$

where L_0 is the likelihood of the data when the coefficient is removed from the model (i.e., set to zero value), whereas L_1 is the likelihood of the data when the coefficient is the estimated value of the coefficient. It is also distributed approximately as a chi-squared probability distribution with one degree of freedom.

Nonlinear regression

Nonlinear regression, `nls()` function, may be a better choice. It can be implemented as follows:

```
attach(yuan)
logisticModel <-nls(cumFreq~DD/(1+exp(-(CC+bb*Months))), start=list(DD=1,CC=0.2,bb=.5
summary(logisticModel)

Formula: yuan$cumFreq ~ DD/(1 + exp(-(CC + bb * yuan$Months)))
```

```
Parameters:
      Estimate   Std. Error    t value    Pr(>|t|)
DD    1.038504     0.014471      71.77    < 2e-16 ***
CC   -4.626982     0.175109     -26.42   5.29e-12 ***
bb    0.206899     0.008777      23.57   2.03e-11 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01908 on 12 degrees of freedom

Number of iterations to convergence: 11
Achieved convergence tolerance: 0.000006909
```

Get fit statistics:

```
AIC(logisticModel)
[1] -71.54679
```

Because AIC for the nonlinear model much smaller (more negative) than AIC for logistic model, we may be tempted to judge fit of the nonlinear regression as best. However, this comparison of models is not valid because the Y variables are different between the two models and the fit families are different. One option is to evaluate fit of models by plots of residuals (see 17.7 – Regression model fit).

Questions

[pending]

Data set

Months	freq	cumFreq
0	0	0
3	0.01063264221159	0.01063264221159
6	0.017012227538543	0.027644869750133
9	0.045188729399256	0.072833599149389
12	0.064327485380117	0.137161084529506
15	0.064859117490697	0.202020202020202
18	0.097820308346624	0.299840510366826
21	0.118553960659224	0.41839447102605
24	0.171185539606592	0.589580010632642
27	0.162147793726741	0.751727804359383
30	0.137161084529506	0.888888888888889
33	0.069643806485912	0.958532695374801
36	0.024455077086656	0.982987772461457
39	0.011695906432749	0.994683678894205
42	0.005316321105795	1

18.4: Generalized Linear Squares

Introduction

Draft

With access to powerful computers and better algorithms, we can move past the classical ANOVA and ordinary least squares approaches to linear models. We have discussed general linear models, but here we introduce **generalized linear models, GLM**. What follows is just a brief foray; for more — and better! discussion, see Zuur et al (2009).

Model variances

Data from Corn and Hiesey (1973) ohia.RData

```
> head(ohia)
   Site  Height   Width
1   M-1 12.5567 19.1264
2   M-1 13.2019 13.1547
3   M-1  8.0699 16.0320
4   M-1  6.0952 22.8586
5   M-1 11.3879 11.0105
6   M-1 12.2242 21.8102
```

ignore the variance issue

```
> AnovaModel.1 <- aov(Height ~ Site, data = ohia); summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value      Pr(>F)
Site         2   4070  2034.8   22.63 0.000000131 ***
Residuals   47   4227    89.9
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternatively, use `gls()` . Default fits by **restricted maximum likelihood**, REML. That is, it's the likelihood of linear combinations of the original data.

```
>model.aov.1 <- gls(Height ~ Site, data = ohia)

Generalized least squares fit by REML
Model: Height ~ Site
Data: ohia
     AIC      BIC    logLik
361.1312 368.5318 -176.5656

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 15.313745  2.120550 7.221591  0.0000
Site[T.M-2] 19.261000  2.998911 6.422666  0.0000
Site[T.M-3]  2.924215  3.672900 0.796160  0.4299

Correlation:
            (Intr) S[T.M-2
```

```
Site[T.M-2] -0.707
Site[T.M-3] -0.577 0.408

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.9832938 -0.5020880 -0.1850871 0.5017636 3.0850635

Residual standard error: 9.483388
Degrees of freedom: 50 total; 47 residual
```

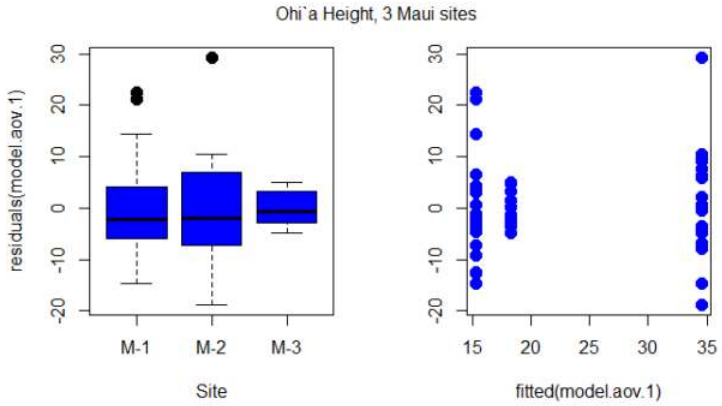


Figure 18.4.1: Box plot of residuals from GLS model by elevation site predictors (left) and scatterplot of residuals by fitted values from GLS model (right).

Code for the plot in Figure 18.4.1:

```
par(mfrow = c(1, 2))
plot(residuals(model.aov.1) ~ Site, pch=19, cex=1.5,col=”blue”, data = ohia)
plot(residuals(model.aov.1) ~ fitted(model.aov.1), pch=19, cex=1.5, col=”blue”, ylab=
mtext(“ANOVA Ohi`a Height, 3 Maui sites “, side = 3, line = -3, outer = TRUE)
```

test equal variances, Height

```
> leveneTest(Height ~ Site, data=ohia, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value  Pr(>F)
group  2  2.1663  0.1259
      47
```

We would conclude no significant departures from equal variances.

```
> bartlett.test(Height ~ Site, data=ohia)

Bartlett test of homogeneity of variances

data:  Height by Site
Bartlett's K-squared = 10.373, df = 2, p-value = 0.005592
```

Bartlett’s test is sensitive to deviations from normality.

Include variances as part of model

```
> model.aov.3 <- gls(Height ~ Site, data = ohia, weights = varIdent(form = ~1|Site));
Generalized least squares fit by REML
Model: Height ~ Site
Data: ohia
    AIC      BIC    logLik
354.421 365.5219 -171.2105
```

`varIdent` permits variances for each group to vary. Results from R continue below.

```
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | Site
 Parameter estimates:
      M-1       M-2       M-3
1.0000000 1.0396880 0.3471771
```

We see here that comparisons were carried out versus the M-1 site.

```
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 15.313745  2.280931 6.713812  0.0000
Site[T.M-2] 19.261000  3.290358 5.853770  0.0000
Site[T.M-3]  2.924215  2.541027 1.150800  0.2556
```

Marginal differences between M-1 and M-2 for height were significantly different, but not between the M-1 and M-3 site.

```
Correlation:
            (Intr) S[T.M-2
Site[T.M-2] -0.693
Site[T.M-3] -0.898 0.622

Standardized residuals:
       Min         Q1        Med        Q3       Max
-1.7734556 -0.6909962 -0.2108834 0.5801370 2.7586550

Residual standard error: 10.20064
Degrees of freedom: 50 total; 47 residual
```

Test the models

```
> anova(model.aov.1, model.aov.3)
            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.aov.1     1  4 361.1312 368.5318 -176.5656
model.aov.3     2  6 354.4210 365.5219 -171.2105 1 vs 2  10.7102 0.0047
```

Although additional degrees of freedom are required, note that this model (`model.aov.3`) has higher (better!) **log likelihood** (-171.21) than `model.aov.1` , the gls model lacking a fit for different variances (-176.57). Introduce a test of the hypothesis that the two models are equal by comparing the log (natural) likelihoods, the **log likelihood ratio test, LRT**.

$$LRT = -2 \cdot \ln\left(\frac{LL\ model_{aov.1}}{LL\ model_{aov.3}}\right) = -2 \cdot \ln[(LL\ model_{aov.1}) - (LL\ model_{aov.3})]$$

The LRT follows a chi-square distribution (per **Wilk's theorem**). If there was no advantage to fitting for unequal variances, then the model fit would not be improved and p-value of the LRT would not be less than 5%.

Conclusion

You can see why this approach, modeling versus separate test of assumptions would be the preferred way to go. We get a better fitting model, cf discussion in

Another example, same data set.

ignore variances, Width

```
model.aov.2 <- gls(Width ~ Site, data = ohia); summary(model.aov.2)
```

Figure 2.

```
par(mfrow = c(1, 2))
plot(residuals(model.aov.2) ~ Site, pch=19, cex=1.5,col="blue", data = ohia)
plot(residuals(model.aov.2) ~ fitted(model.aov.2), pch=19, cex=1.5, col="blue", ylab=
mtext("ANOVA Ohi`a Width 3 Maui sites ", side = 3, line = -3, outer = TRUE)
```

test equal variances, Width

```
Tapply(Width ~ Site, var, na.action=na.omit, data=ohia) # variances by group
leveneTest(Width ~ Site, data=ohia, center="median")
Tapply(Width ~ Site, var, na.action=na.omit, data=ohia) # variances by group
bartlett.test(Width ~ Site, data=ohia)
```

model the variances, Height

```
library(nlme)
model.aov.3 <- gls(Height ~ Site, data = ohia, weights = varIdent(form = ~1|Site)); s
par(mfrow = c(1, 2))
plot(residuals(model.aov.3) ~ Site, pch=19, cex=1.5,col="red", data = ohia)
plot(residuals(model.aov.3) ~ fitted(model.aov.3), pch=19, cex=1.5, col="red", ylab="
mtext("GLS Ohi`a Height 3 Maui sites ", side = 3, line = -3, outer = TRUE)
```

Test the models

```
anova(model.aov.1, model.aov.3)
```

model the variances, Width

```
model.aov.4 <- gls(Width ~ Site, data = ohia, weights = varIdent(form = ~1|Site)); su
par(mfrow = c(1, 2))
plot(residuals(model.aov.4) ~ Site, pch=19, cex=1.5,col="red", data = ohia)
```

```
plot(residuals(model.aov.4) ~ fitted(model.aov.4), pch=19, cex=1.5, col="red", ylab="
mtext("GLS Ohi`a Width 3 Maui sites ", side = 3, line = -3, outer = TRUE)
```

Test the models

```
anova(model.aov.2, model.aov.4)
```

Model correlated residuals

[pending]

Questions

[pending]

18.5: Selecting the best model

Introduction

This is a long entry in our textbook, with many topics to cover. We discuss aspects of **model fitting**, from why model fitting is done to how to do it and what statistics are available to help us decide on the **best model**. Model selection very much depends on what the intent of the study is. For example, if the purpose of **model building** is to provide the best description of the data, then in general one should prefer the **full** (also called the **saturated**) **model**. On the other hand, if the purpose of model building is to make a predictive statistical model, then a **reduced model** may prove to be a better choice. The text here deals mostly with the later context of model selection, finding a justified reduced model.

From Full model to best Subset model

Model building is an essential part of being a scientist. As scientists, we seek models that explain as much of the variability about a phenomenon as possible, but yet remain simple enough to be of practical use.

Having just completed the introduction to multiple regression, we now move to the idea of how to pick best models.

We distinguish between a full model, which includes as many variables (predictors, factors) as the regression function can work with, returning interpretable, if not always statistically significant output, and a saturated model.

The saturated model is the one that includes all possible predictors, factors, and interactions in your experiment. In well-behaved data sets, the full model and the saturated model will be the same model. However, they need not be the same model. For example, if two predictor variables are highly **collinear**, then you may return an error in regression fitting.

For those of you working with **meta-analysis** problems, you are unlikely to be able to run a saturated model because some level of a key factor are not available in all or at least most of the papers. Thus, in order to get the model to run, you start dropping factors, or you start nesting factors. If you were unable to get more things in the model, then this is your “full” model. Technically we wouldn’t call it saturated because there were other factors, they just didn’t have enough data to work with or they were essentially the same as something else in the model.

Identify the model that does run to completion as your full model and proceed to assess model fit criteria for that model, and all reduced models thereafter.

In R (Rcmdr) you know you have found the full model when the output lacks “NA” strings (**missing values**) in the output. Use the full model to report the values for each coefficient, i.e., conducting the inferential statistics.

Get the estimates directly from the output from running the regression function. You can tell if the effect is positive (look at the estimate for sample — it is positive) so you can say — more samples, greater likelihood to see more cases of cancer.

Remember, the experimental units are the papers themselves, so studies with larger numbers of subjects are going to find more cases of diabetes. We would worry big time with your project if we did not see statistically significant and positive effects for sample size.

For illustration, here’s an example output following a run with the linear model function on an experimental data set.

The variables were

BMI = Dependent variable, continuous]

Age = Independent variable, continuous

CalsPDay = Independent variable, continuous

CholPDay = Independent variable, continuous

Sex = Independent variable, categorical

Smoke = Independent variable, categorical

```
lm(formula = BMI ~ Age + CalsPDay + CholPDay + Sex + Smoke +
Sex:Smoke, data = BMI) Residuals:
Min 1Q Median 3Q Max
```

```
-9.9685 -3.3766 -0.6609 2.5090 22.3482

Coefficients:
                          Estimate   Std. Error   t value    Pr(>F)
(Intercept)             25.9351297    3.7205047     6.971   1.708e-09 ***
Age                      0.890
CalsPDay                -0.0005757    0.0009882    -0.583    0.562
 CholPDay                0.0103521    0.0060722     1.705    0.093 .
 Sex[T.M]               -0.8529925    2.2209045    -0.384    0.702
 Smoke[T.Yes]           -1.1670159    1.9134734    -0.610    0.544
Sex[T.M]:Smoke[T.Yes]    0.9261469    2.8510680     0.325    0.746
```

The Y-variable was BMI, and the predictor variables included gender (male, female), smokers (yes, no), and the interaction, plus two measures of diet quality (calories per day and amount of cholesterol).

Question. Write out the equation in symbol form.

We see that none of the factors or covariates were statistically significant, so I wouldn't go on and on about positive or negative.

But, for didactic purposes here, imagine the P-value for `CholPDay` was less than 0.05 (and therefore statistically significant). We report the value (0.0103521 → I would round to 0.01), and note that those who had more cholesterol in their diet per day, those individuals tended to have higher BMI (e.g., the sign of the coefficient — and I related the coefficient back to the most important thing about your study — the biological interpretation).

Now's a good time to be clear about HOW you report statistical results. DO NOT SIMPLY COPY AND PASTE EVERYTHING into your report. Now, for the estimates above, you would report everything, but not all of the figures. Here's how the output should be reported in your Project paper.

Table 18.5.1. Coefficients from full model.

```
                     Estimate      SE       t   P-value
Intercept              25.935   3.721    6.97  < 0.0001
Age                    -0.007   0.051   -0.14    0.8892
Calories/Day           -0.001   0.001   -0.58    0.5622
Cholesterol/Day         0.010   0.006    1.71    0.0929
Sex                    -0.853   2.221   -0.38    0.7021
Smoke                  -1.167   1.915   -0.61    0.5440
Interaction Smoke:Sex   0.926   2.851    0.33    0.7463
```

Looks better, doesn't it?

Once you have the full model, use this model for the inferential statistics. Use the significance tests of each parameter in the model from the corresponding ANOVA table. Now, where is the ANOVA table? Remember, right after running the linear regression,

Rcmdr: Models → Hypothesis testing → ANOVA tables

Accept the default (**partial marginality**), and, Boom! Out pops the ANOVA table you should be familiar with.

From the ANOVA table you will tell me whether a Factor is significant or not. You report the ANOVA table in your paper. You describe it.

Now, the next step is to decide what is the best model. It then guides you to the next step which is to decide whether a better model (fewer parameters, Occam's razor) can be found. Identify the parameter from the ANOVA table with the highest P-value and remove it from the model when you run the regression again. Repeat the steps above, return the **ANOVA table**, checking the estimates and P-values, until you have a model with only statistically significant parameters.

Find the best model

Output from R follows:

```
Anova(LinearModel.1, type="II")
Anova Table (Type II tests)

Response: BMI
             Sum Sq   Df     F value        P
Age            0.62    1      0.0196    0.890
CalsPDay      10.79    1      0.3394    0.562
CholPDay      92.37    1      2.9065    0.093
Sex            1.52    1      0.0478    0.828
Smoke          8.84    1      0.2782    0.600
Sex:Smoke      3.35    1      0.1055    0.746
Residuals   2129.34   67
```

This is my full model and I would start anticipating the need to reduce my model because none of the factors are statistically significant. By the criterion that simple models are better, I would proceed first to drop the interaction. See below for more on selecting the best models.

But first, I want to take up an important point about your models that you may not have had a chance to think about. The order of entry of parameters in your model can effect the significance and value of the estimates themselves. The order of parameter model entry above can be read top to bottom. Age was first, followed in sequence by CalsPDay, CholPDay, and so on. By convention, enter the covariates first (the ratio-scale predictors), that's what I did above.

Here's the output from a model in which I used a different order of parameters.

```
Anova(LinearModel.2, type="II")
Anova Table (Type II tests)

Response: BMI
              Sum Sq    Df    F value     Pr(>F)
Sex             1.52     1     0.0478    0.82764
Smoke           8.84     1     0.2782    0.59964
Age             0.62     1     0.0196    0.88918
CalsPDay       10.79     1     0.3394    0.56215
CholPDay       92.37     1     2.9065    0.09286 .
Sex:Smoke       3.35     1     0.1055    0.74631
Residuals    2129.34    67
```

The output is the same!!! So why did I give you a warning about parameter order? Run the ANOVA table summary command again, but this time select **Type III type of test**, i.e., ignore marginality.

```
> Anova(LinearModel.2, type="III")
Anova Table (Type III tests)

Response: BMI
               Sum Sq    Df    F value       Pr(>F)
(Intercept)   1544.34     1    48.5929   1.708e-09 ***
Sex              4.69     1     0.1475     0.70214
```

```
Smoke           11.82     1     0.3720     0.54400
Age              0.62     1     0.0196     0.88918
CalsPDay        10.79     1     0.3394     0.56215
CholPDay        92.37     1     2.9065     0.09286 .
Sex:Smoke        3.35     1     0.1055     0.74631
Residuals     2129.34    67
```

The output has changed — and in fact it now reports the significance test of the intercept. This output is the same as the output from the linear model. Try again, this time selecting Type I, sequential:

```
> anova(LinearModel.2)
Analysis of Variance Table

Response: BMI
           Df   Sum Sq    Mean Sq  F value    Pr(>F)
Sex         1     0.68      0.681   0.0214   0.88402
Smoke       1     2.82      2.816   0.0886   0.76690
Age         1     3.44      3.436   0.1081   0.74333
CalsPDay    1     2.27      2.272   0.0715   0.78998
CholPDay    1    96.30     96.299   3.0301   0.08633
Sex:Smoke   1     3.35      3.354   0.1055   0.74631
Residuals  67  2129.34     31.781
```

Here, we see the effect of order. So, as we are working to learn all of the issues of statistics and in particular mode fitting, I have purposefully restricted you to **Type II analyses** — obeying **marginality** correctly handles most issues about **order of entry**.

Status check

Where are we??? Recall that the purpose of all of this effort is to find the best supported model. The question we are working on is whether the full (saturated) model is the best model or if a reduced model can be supported.

We go back to my first full model output from ANOVA.

Model 1:

```
Anova(LinearModel.1, type="II")
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value    Pr(>F)
Age           0.62    1    0.0196   0.88918
CalsPDay     10.79    1    0.3394   0.56215
CholPDay     92.37    1    2.9065   0.09286 .
Sex           1.52    1    0.0478   0.82764
Smoke         8.84    1    0.2782   0.59964
Sex:Smoke     3.35    1    0.1055   0.74631
Residuals  2129.34   67
```

We have two factors (Sex, Smoke), three covariates (Age, CalsPDay, CholPDay), and one two-way interaction (Sex:Smoke). We would write our full model then as

$$BMI \sim Age + CalsPDay + CholPDay + Sex + Smoke + Sex : Smoke$$

Get and save in your output the ANVOA table for this Full model. Proceed to test a series of nested reduced models. Start by dropping the interaction terms, consistent with our Occam's razor approach.

Model 2:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value    Pr(>F)
Age           0.63    1    0.0201   0.88760
CalsPDay     10.45    1    0.3331   0.56572
CholPDay     96.30    1    3.0704   0.08424 .
Sex           1.52    1    0.0484   0.82650
Smoke         8.84    1    0.2819   0.59720
Residuals  2132.69   68
```

Next, reduce by identifying Factors or Predictors with the highest P-values. Looks like "Age" is next.

Model 3:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value   Pr(>F)
CalsPDay      9.87    1    0.3194  0.57381
CholPDay     97.78    1    3.1627  0.07974 .
Sex           2.55    1    0.0824  0.77488
Smoke         8.61    1    0.2786  0.59934
Residuals  2133.32   69
```

Next up, drop the "Sex" parameter.

Model 4:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df  F value    Pr(>F)
CalsPDay     10.45    1   0.3424   0.56032
CholPDay     96.12    1   3.1502   0.08027 .
Smoke        12.42    1   0.4070   0.52557
Residuals  2135.87   70
```

Next? You would select CalsPDay, right?

Model 5:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value    Pr(>F)
```

```
CholPDay     90.24    1    2.9852   0.08837 .
Smoke        13.95    1    0.4613   0.49923
Residuals  2146.32   71
```

And finally, we remove the “Smoke” factor.

Model 6:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value   Pr(>F)
CholPDay     77.93    1    2.5974   0.1114
Residuals  2160.26   72
```

Oops!

What happened to Model 6? Nothing remains significant? Panic? What is the point??? Arggh, Dr Dohm…!!!

Easy there…. Take a deep breath, and guess what? Your best model needs to have significant parameters in it, right? Your best fit model then is Model 5. And that model will be your candidate for best fit as we proceed to complete our model building.

$$BMI \sim CholPDay + Smoke$$

Now we proceed to gain some support evidence for our candidate best model. We are going to use an information criterion approach.

Use a fit criterion for determining model fit

To help us evaluate evidence in favor of one model over another there are a number of statistics one may calculate to provide a single number for each model for comparison purposes. The criteria model evaluators available to us include **Mallow’s C_{p}\)**, **adjusted R^2**, Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to select best model.

We already introduced the **coefficient of determination** R^2 as a measure of fit – in general we favor models with larger values of R^2. However, values of R^2 will always be larger for models with more parameters. Thus, the other evaluators attempt to adjust for the parameters in the model and how they contribute to increased model fit. For illustrative purposes we will use Mallow’s C_p. The equation for Mallow’s C_p in linear regression is

$$C_p = \frac{Reduced\ SS_{residual}}{Full\ MS_{residual}} - [n - 2(p+1)]$$

where p is the number of parameters in the model. Mallow’s C_p is thus equal to the number of parameters in the model plus an additional amount due to lack of fit of the model (i.e., large residuals). All else being equal we favor the model in which the Cp is close to the number of parameters in the model.

In Rcmdr, select **Models** → **Subset model selection …** (Fig. 18.5.1)

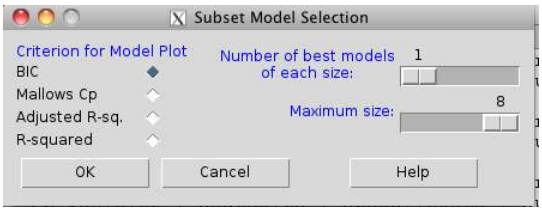


Figure 18.5.1: Rcmdr popup menu, Subset model selection…

From the menu, select the criterion and how many models to return. The function returns a graph that can be used to interpret which model is best given the selection criterion used. Below is an example (although for a different data set!) for Mallow’s C_p (Fig. 18.5.2).

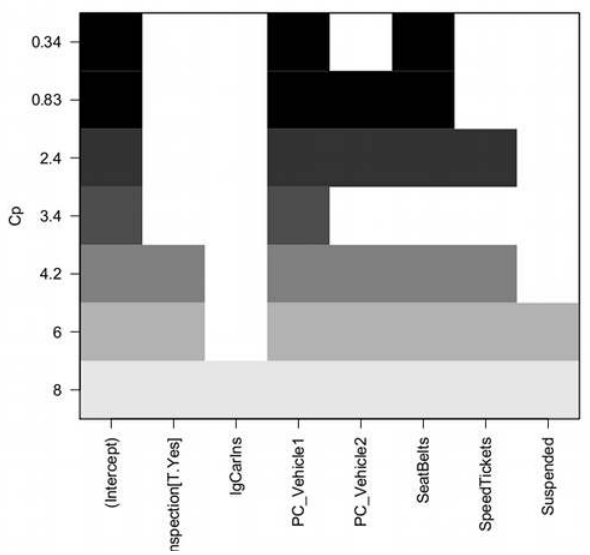


Figure 18.5.2: Mallow's C_p plot

Let's break down the plot. First define the axes. The vertical axis is the range of values for the C_p calculated for each model. The horizontal axis is categorical and reads from left to right: Intercept, Inspection[T.Yes], etc., up to Suspended. Looking into the graph itself we see horizontal bars — the extent of shading indicates which model corresponds to the C_p value. For example, the lowest bar, which is associated with the C_p value of 8, extends all the way to the right of the graph. This says that the model evaluated included all of the variables and therefore was the saturated or full model. The next bar from the bottom of the graph is missing only one block (lgCarIns), which tells us the C_p value 6 corresponds to a reduced model, and so forth.

Cross-validation

Once you have identified your Best Fit model, then, your proceed to run the diagnostics plots. For the rest of the discussion we return to our first example.

Rcmdr: **Models → Graphs → Basic diagnostic plots**.

We'll just concern ourselves with the first row of plots (Fig. 18.5.3).

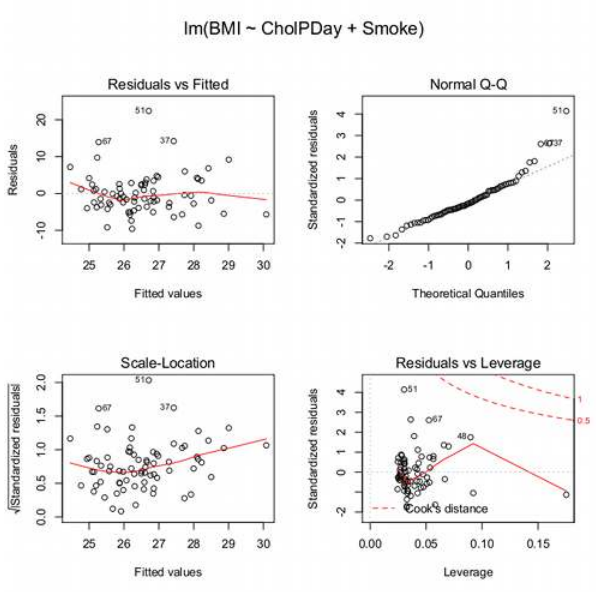


Figure 18.5.3: Diagnostic plots.

The left one shows the residuals versus the predicted values — if you see a trend here, the assumption of linearity has been violated. The second plot is a test of the assumption of normality of the residuals. Interpret them (residuals OK, Residuals normally distributed? Yes/No), and you're done. Here, I would say I see no real trend in the residuals vs. fitted plot, so assumption of linear

fit is OK. For normality, there is a tailing off at the larger values of residuals, which might be of some concern (and I would start thinking about possible leverage problems), but nothing dramatic. I would conclude that our Model 5 is a good fitting model and one that could be used to make predictions.

Now, if you think a moment, you should identify a logical problem. We used the same data to "check" the model fit as we did to make the model in the first place. In particular if the model is intended to make predictions it would be advisable to check the performance of the model (e.g., does it make reasonable predictions?) by supplying new data, data not used to construct the model, into the model. If new data are not available, one acceptable practice is to divide the full data set into at least two subsets, one used to develop the model (sometimes called the calibration or training dataset) and the other used to test the model. The benefits of cross-validation include testing for influence points, over fitting of model parameters, and a reality check on the predictions generated from the model.

Questions

[pending]

18.6: Compare two linear models

Rcmdr (R) provides a very useful tool to compare models. Now, you can compare any two models, but this would be a poor strategy. Use this tool to perform in effect a stepwise test by hand. As one of the models, select for example the saturated model, then for the second model, select one in which you drop one model factor. In the example below, I dropped the two-way interaction from the saturated model (a logistic regression model, actually):

The model was

$$Type.IIdiabetes\,Treatment + Samples + Gender + BMI + Age + Gender : Treatment$$

where Type.II diabetes is a binomial (Yes,No) dependent variable and Treatment and Gender were categorical factors. The ANOVA table is shown below.

```
Anova(GLM.1, type="II", test="LR")
Analysis of Deviance Table (Type II tests)
Response: Type.II
               LR Chisq  Df Pr(>Chisq)
Treatment         0.266  7    0.9999
Samples          38.880  1    4.508e-10 ***
Gender            0.671  1    0.4127
BMI               2.259  1    0.1329
Age               2.064  1    0.1508
Treatment:Gender  1.803  1    0.1794
```

From this output we see that there are a number of terms that are not significant $(P < 0.05)$, but with one exception (Treatment) they seem to contribute to the total variation (P values are between 0.13 and 0.4). So, we conclude that the saturated model is not the best fit model, and proceed to evaluate alternative models in search of the best one.

As a matter of practice I first drop the interaction term. Here's the ANOVA table for the second model, now without the interaction:

```
Anova(GLM.1, type="II", test="LR")
Analysis of Deviance Table (Type II tests)
Response: Type.II
             LR Chisq Df Pr(>Chisq)
Treatment       0.266  7    0.9999
Samples        37.086  1    1.13e-09 ***
Gender          0.671  1    0.4127
BMI             2.017  1    0.1556
Age             1.794  1    0.1804
```

Both models look about the same. Which one is best? We now wish to know if dropping the interaction harms the model in any way. We will use the AIC (Akaike Information Criterion) to evaluate the models. AIC provides a way to assess which among a set of nested models is better. The preferred model is the one with the lowest AIC value.

To access the AIC calculation, just enter the script AIC(model name), where model name refers to one of the models you wish to evaluate (e.g., GLM.1), then submit the code

```
AIC(GLM.1)
50.65518
```

```
AIC(GLM.2)
50.45793
```

Thus, we prefer the second model (GLM.2) because the AIC is lower.

AIC does not provide a statistical test of model fit. To access the model comparison tool, simply select

Models → Hypothesis tests → Compare two models…

and the following screen will appear.

Figure 18.6.1: Compare Models menu in R Commander.

Select the two models to compare (in this case, GLM1 and GLM2), then press OK button. R output:

```
anova(GLM.1, GLM.2, test="Chisq")
Analysis of Deviance Table
Model 1: Type.II ~ Treatment + Samples + Gender + BMI + Age + Gender:Treatment
Model 2: Type.II ~ Treatment + Samples + Gender + BMI + Age
    Resid. Df  Resid. Dev  Df  Deviance P(>|Chi|)
1          49      24.655
2          50      26.458  -1   -1.8027   0.1794
```

We see that $P > 0.05 (= 0.1794)$, which means the fit of the model is fine if we lose the one term.

Deviance

Those of you working with logistic regressions will see this new term, “deviance.” Deviance is a statistical term relevant to model fitting. Think of it like a chi-square test statistic. The idea is that you compare your fitted model against the data in which the only thing estimate is the intercept. Do the additional components of the model add significantly to the prediction of the original data? If they do, dropping the term will have a significant effect on the model fit and the P-value would be less than 0.05. In this example, we see that dropping the interaction term had little effect on the deviance score and in agreement, the P value is larger than 0.05. It means we can drop the term and the new model lacking the term is in some sense better: fewer predictors, a simpler model.

Questions

[pending]

18.7: References and suggested readings (Ch. 17 and 18)

Bewick, V., Cheek, L., Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care* 9:112-118.

Corn, C. A., & Hiesey, W. M. (1973). Altitudinal ecotypes in Hawaiian Metrosideros. *American Journal of Botany* 60(10): 991-1002.

Faraway, J. J. Practical Regression and ANOVA with R 2002.

Gosner, Kenneth L. (1960). A simplified table for staging anuran embryos and larvae with notes on identification. *Herpetologica* 16 (3): 183–190. link to Wikipedia page

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol*, *13*(3), e1002106.

LaBarber,a M. (1989). Analyzing body size as a factor in ecology and evolution. *Annual Review of Ecology and Systematics* 29:97-117.

McArdle, B. H. (2003). Lines, models, and errors: Regression in the field. *Limnology and Oceanography* 48:1363-1366.

McCullagh, P. 2002. What is a statistical model? *The Annals of Statistics* 30(5): 1225-1310.

Meeker, W. Q., Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *American Statistician* 49:48-53.

O'Brien, R. M. (2007) A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* 41:673-690

Pawitan, Y. (2000). A Reminder of the Fallibility of the Wald Statistic: Likelihood Explanation. *American Statistician* 54:54-56.

Slinker, B. K., & Stanton, G. (2008). A Multiple Linear Regression: Accounting for Multiple Simultaneous Determinants of a Continuous Dependent Variable. *Circulation* 117, 1732-1737

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346. https://doi.org/10.1098/rsos.220346.

Warton, D. I., Wright, I. J., Falster, D. S., Westoby, M. (2006). Bivariate line-fitting methods for allometry. *Biological Reviews* 81: 259-291

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer.

CHAPTER OVERVIEW

19: Distribution-free methods

Introduction

We introduced the concept of permutation tests in our chapter on parameter estimates and statistical error (Chapter 3.3). **Jackknife** and **bootstrapping** are permutation approaches to working with data when the **Central Limit theorem** is unlikely to apply or, rather, we don't wish to make that assumption. The jackknife is a sampling method involving repeatedly sampling from the original data set, but each time leaving one value out. The **estimator**, for example, the sample mean, is calculated for each sample. The repeated estimates from the jackknife approach yield many estimates which, collected, are used to calculate the sample variance. Jackknife estimators tend to be less biased than those from classical asymptotic statistics. Bootstrapping, and not jackknife resampling, is now the preferred permutation approach (add citations).

Bootstrapping

Bootstrapping involves large numbers of **permutations** of the data, which, in short, means we repeatedly take many samples of our data and recalculate our statistics on these sets of sampled data. We obtain statistical significance by comparing our result from the original data against how often results from our permutations on the resampled data sets exceed the originally observed results. By permutation methods, the goal is to avoid the assumptions made by large-sample statistical inference. Since its introduction, "bootstrapping" has been shown to be superior in many cases for statistics of error compared to the standard, classical approach (add citations).

Permutation vs classical NHST approach

There are many advocates for this approach, and, because we have computers now instead of the hand calculators our statistics ancestors used, permutation methods may be the approach you will take in your own work. However, the classical approach has it's strengths; when the conditions, that is, when the assumptions of asymptotic statistics are met by the data, then the classical approaches tend to be less conservative than the permutation methods. By conservative, statisticians mean that a test performs at the level we expect it to. Thus, if the assumptions of classical statistics are met they return the correct answer more often than do the permutation tests.

19.1: Jackknife sampling

Introduction

edits: — under construction —

R packages

There are several R packages one could use. The package `bootstrap` may be the the most general, and includes a jackknife routine suitable for any function. This page demonstrates jackknife estimate of correlation.

Example data set of cars, showing stopping distance by speed of car (scroll down or click here).

install package bootstrap

Jackknife estimates on linear models

These procedures can be done with the bootstrap package, but `lmboot` is a specific package to solve the problem

install package `lmboot`

Example data set, Tadpoles from Chapter 14, copied to end of this page for your convenience (scroll down or click here).

R code

```
jackknife(VO2~Body.mass, data = Tadpoles)
```

R returns two values:

1. `bootEstParam` , which are the jackknife parameter estimates. Each column in the matrix lists the values for a coefficient. For this model, `bootEstParam$[,1]` is the intercept and `bootEstParam$[,2]` is the slope.
2. `origEstParam` , a vector with the original parameter estimates for the model coefficients.

```
$bootEstParam
     (Intercept) Body.mass
[1,]   -660.8403  472.6841
[2,]   -539.5951  430.3990
[3,]   -612.8495  454.5188
[4,]   -512.5914  423.0815
[5,]   -543.1577  434.2789
[6,]   -572.3895  442.9176
[7,]   -613.7873  451.2656
[8,]   -594.0366  446.2571
[9,]   -582.1833  443.5404
[10,]  -598.2244  456.0599
[11,]  -531.3152  415.2467
[12,]  -555.7287  430.5604
[13,]  -726.8522  512.1268

$origEstParam
[,1]
(Intercept) -583.0454
Body.mass 444.9512
```

Get necessary statistics and plots

```
#95% CI slope
quantile(jack.model.1$bootEstParam[,2], probs=c(.025, .975))
```

R returns

```
    2.5%    97.5%
417.5971 500.2940
```

```
#95% CI intercept
quantile(jack.model.1$bootEstParam[,1], probs=c(.025, .975))
```

R returns

```
     2.5%     97.5%
-707.0486 -518.2085
```

Coefficient estimates

Slope

```
#plot the sampling distribution of the slope coefficient
par(mar=c(5,5,5,5)) #setting margins to my preferred values
hist(jack.model.1$bootEstParam[,2], col="blue", main="Jackknife Sampling Distribution
xlab="Slope Estimate")
```

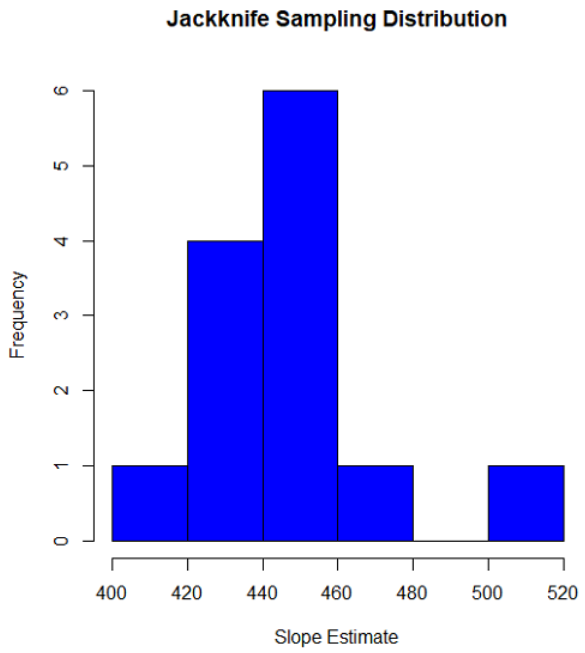


Figure 19.1.1: Histogram of jackknife estimates for slope.

Intercept

```
#95% CI intercept
quantile(jack.model.1$bootEstParam[,1], probs=c(.025, .975))
```

```
par(mar=c(5,5,5,5))
hist(jack.model.1$bootEstParam[,1], col="blue", main="Jackknife Sampling Distribution
xlab="Intercept Estimate")
```

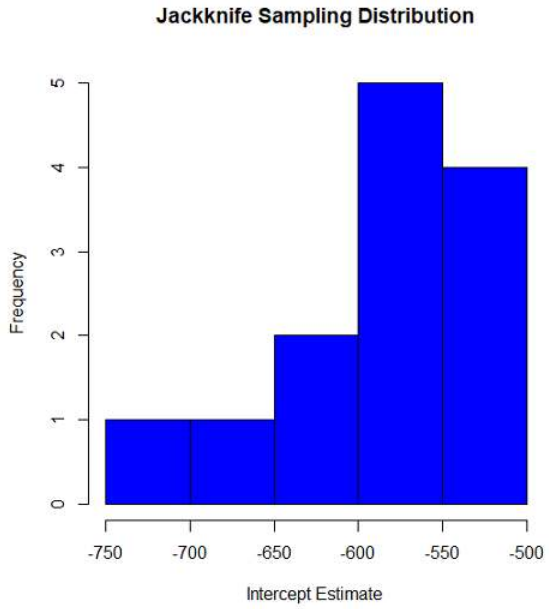


Figure 19.1.2: Histogram of jackknife estimates for intercept.

Questions

edits: pending

Cars data set used in this page

speed	dist
4	2
4	10
7	4
7	22
8	16
9	10
10	18
10	26
10	34
11	17
11	28

12	14
12	20
12	24
12	28
13	26
13	34
13	34
13	46
14	26
14	36
14	60
14	80
15	20
15	26
15	54
16	32
16	40
17	32
17	40
17	50
18	42
18	56
18	76
18	84
19	36
19	46

19	68
20	32
20	48
20	52
20	56
20	64
22	66
23	54
24	70
24	92
24	93
24	120
25	85

Tadpole data set used in this page (sorted)

Gosner	Body mass	VO2
I	1.76	109.41
I	1.88	329.06
I	1.95	82.35
I	2.13	198
I	2.26	607.7
II	2.28	362.71
II	2.35	556.6
II	2.62	612.93
II	2.77	514.02

Gosner	Body mass	VO2
II	2.97	961.01
II	3.14	892.41
II	3.79	976.97
NA	1.46	170.91

19.2: Bootstrap sampling

Introduction

Bootstrapping is a general approach to estimation or statistical inference that utilizes random **sampling with replacement** (Kulesa et al. 2015). In classic frequentist approach, a sample is drawn at random from the population and assumptions about the population distribution are made in order to conduct statistical inference. By resampling with replacement from the sample many times, the **bootstrap samples** can be viewed as if we drew from the population many times without invoking a theoretical distribution. A clear advantage of the bootstrap is that it allows estimation of confidence intervals without assuming a particular theoretical distribution and thus avoids the burden of repeating the experiment.

Base install of R includes the `boot` package. The boot package allows R users to work with most functions, and many authors have provided helpful packages. I highlight a couple packages

install packages `lmboot, confintr`

Example data set, Tadpoles from Chapter 14, copied to end of this page for convenience (scroll down or click here).

Bootstrapped 95% Confidence interval of population mean

Recall the classic frequentist (large-sample) approach to confidence interval estimates of mean using R:

```
x = round(mean(Tadpole$Body.mass),2); x
n = length(Tadpole$Body.mass); n
s = sd(Tadpole$Body.mass); s
error = qt(0.975,df=n-1)*(s/sqrt(n)); error
lower_ci = round(x-error,3)
upper_ci = round(x+error,3)
paste("95% CI of ", x, " between:", lower_ci, "&", upper_ci)
```

Output results are

```
> n = length(Tadpole$Body.mass); n
[1] 13
> s = sd(Tadpole$Body.mass); s
[1] 0.6366207
> error = qt(0.975,df=n-1)*(s/sqrt(n)); error
[1] 0.384706
> paste("95% CI of ", x, " between:", lower_ci, "&", upper_ci)
[1] "95% CI of 2.41 between: 2.025 & 2.795"
```

We used the **t-distribution** because both μ, the population mean, and σ, the population standard deviation, were unknown. Thus, 95 out of 100 **confidence intervals** would be expected to include the true value.

Bootstrap equivalent:

```
library(confintr)
ci_mean(Tadpole$Body.mass, type=c("bootstrap"), boot_type=c("stud"), R=999, probs=c(0
```

Output results are

```
Two-sided 95% bootstrap confidence interval for the population mean based on 999 boot
and the student method
```

```
Sample estimate: 2.412308
Confidence interval:
    2.5%    97.5%
2.075808 2.880144
```

where `stud` is short for student t distribution (another common option is the percentile method — replace `stud` with `perc`), R = 999 directs the function to resample 999 times. We set seed=1 to initialize the **pseudorandom number generator** so that if we run the command again, we would get the same result. Any integer number can be used. For example, I set `seed = 1` for the output below:

```
Confidence interval:
    2.5%    97.5%
2.075808 2.880144
```

Compare it to repeated runs without initializing the pseudorandom number generator:

```
Confidence interval:
    2.5%    97.5%
2.067558 2.934055
```

and again

```
Confidence interval:
    2.5%    97.5%
2.067616 2.863158
```

Note that the classic confidence interval is narrower than the bootstrap estimate, in part because of the small sample size (i.e., not as accurate, does not actually achieve the nominal 95% coverage). Which to use? The sample size was small, just 13 tadpoles. Bootstrap samples were drawn from the original data set, thus it cannot make a small study more robust. The 999 samples can be thought as estimating the sampling distribution. If the assumptions of the t-distribution hold, then the classic approach would be preferred. For the `Tadpole` data set, `Body.mass` was approximately normally distributed (**Anderson-Darling test** = 0.21179, p-value = 0.8163). For cases where assumption of a particular distribution is unwarranted (e.g., what is the appropriate distribution when we compare medians among samples?), bootstrap may be preferred (and for small data sets, percentile bootstrap may be better). To complete the analysis, percentile bootstrap estimate of confidence interval are presented.

The R code

```
ci_mean(Tadpole$Body.mass, type=c("bootstrap"), boot_type=c("perc"), R=999, probs=c(0
```

and the output

Two-sided 95% bootstrap confidence interval for the population mean, based on 999 bootstrap replications and the percent method:

```
Sample estimate: 2.412308
Confidence interval:
    2.5%    97.5%
2.076923 2.749231
```

In this case, the bootstrap percentile confidence interval is narrower than the frequentist approach.

Model coefficients by bootstrap

R code

Enter the model, then set `B`, the number of samples with replacement.

```
myBoot <- residual.boot(VO2~Body.mass, B = 1000, data = Tadpoles)
```

R returns two values:

1. `bootEstParam`, which are the bootstrap parameter estimates. Each column in the matrix lists the values for a coefficient. For this model, `bootEstParam$[,1]` is the intercept and `bootEstParam$[,2]` is the slope.
2. `origEstParam`, a vector with the original parameter estimates for the model coefficients.
3. `seed`, numerical value for the seed; use seed number to get reproducible results. If you don't specify the seed, then seed is set to pick any random number.

While you can list the `$bootEstParam`, not advisable because it will be a list of 1000 numbers (the value set with B)!

Get necessary statistics and plots

```
#95% CI slope
quantile(myBoot$bootEstParam[,2], probs=c(.025, .975))
```

R returns

```
    2.5%    97.5%
335.0000 562.6228
```

```
#95% CI intercept
quantile(myBoot$bootEstParam[,1], probs=c(.025, .975))
```

R returns

```
     2.5%     97.5%
-881.3893 -310.8209
```

Slope

```
#plot the sampling distribution of the slope coefficient
par(mar=c(5,5,5,5)) #setting margins to my preferred values
hist(myBoot$bootEstParam[,2], col="blue", main="Bootstrap Sampling Distribution",
xlab="Slope Estimate")
```

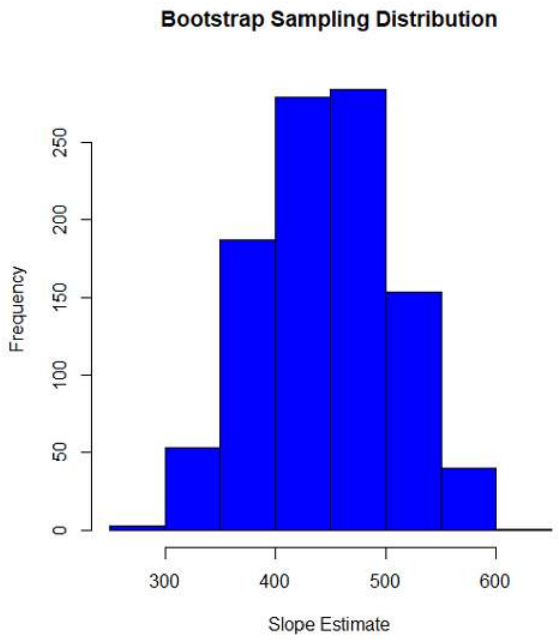


Figure 19.2.1: Histogram of bootstrap estimates for slope.

Intercept

```
#95% CI intercept
quantile(myBoot$bootEstParam[,1], probs=c(.025, .975))
par(mar=c(5,5,5,5))
hist(myBoot$bootEstParam[,1], col="blue", main="Bootstrap Sampling Distribution",
xlab="Intercept Estimate")
```

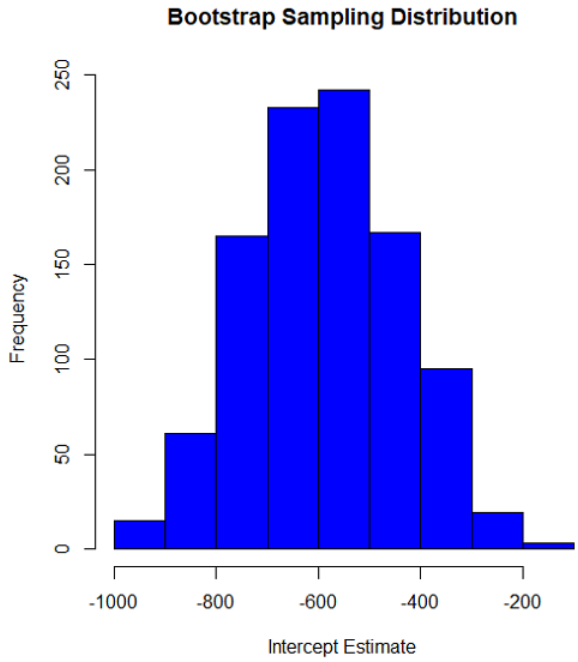


Figure 19.2.2: Histogram of bootstrap estimates for intercept.

Questions

edits: pending

Data set used in this page (sorted)

Gosner	Body mass	VO2
I	1.76	109.41
I	1.88	329.06
I	1.95	82.35
I	2.13	198
I	2.26	607.7
II	2.28	362.71
II	2.35	556.6
II	2.62	612.93
II	2.77	514.02
II	2.97	961.01
II	3.14	892.41
II	3.79	976.97
NA	1.46	170.91

19.3: Monte Carlo methods

edits: — under construction —

Introduction

Statistical methods that employ Monte Carlo methods use repeated random sampling to estimate properties of a frequency distribution. These distributions may be well-known, e.g., gamma-distribution, normal distribution, or t-distribution. The simulation is based on generation of a set of random numbers on the **open interval** $(0, 1)$ — the set of real numbers between zero and one (all numbers greater than 0 and less than 1).

> **Note:**
>
> If the set included 0 and 1, then it would be called a **closed set**, i.e., the set includes the boundary points zero and one.

The Markov chain Monte Carlo (MCMC) sampling approach can be used to solve large scale problems. The Markov chain refers to how the sample is drawn from a specified probability distribution. It can be drawn by discrete time steps (DTMC) or by a continuous process (CTMC). The Markov process is "memoryless:" predictions of future events are derived solely from their present state — the future and past states are independent.

Gibbs sampling is a common MCMC algorithm.

R code

R's uniform generator is `runif` function. Examples of the samples generated over different values (100, 1000, 10000, 100000) with output displayed as histograms (Fig. 1). Note that as sample size increases, the simulated distributions resemble more and more the uniform distribution. Use `set.seed()` to reproduce the same set and sequence of numbers

```
require(RcmdrMisc)
par(mfrow = c(2, 2))
myUniformH <- data.frame(runif(100))
with(myUniformH, Hist(runif.100., scale="frequency", ylim=c(0,20), breaks="Sturges",
myUniform1K <- data.frame(runif(1000))
with(myUniform1K, Hist(runif.1000., scale="frequency", ylim=c(0,150), breaks="Sturges
myUniform10K <- data.frame(runif(10000))
with(myUniform10K, Hist(runif.10000., scale="frequency", ylim=c(0,600), breaks="Sturg
myUniform100K <- data.frame(runif(100000))
with(myUniform100K, Hist(runif.100000., scale="frequency", ylim=c(0,5000),breaks="Stu
#reset par()
dev.off()
```

> **Note:**
>
> Yes, a nice repeating function would be more elegant code, but we move on. As a suggestion, you should create one! Use `sapply()` or a basic for loop.

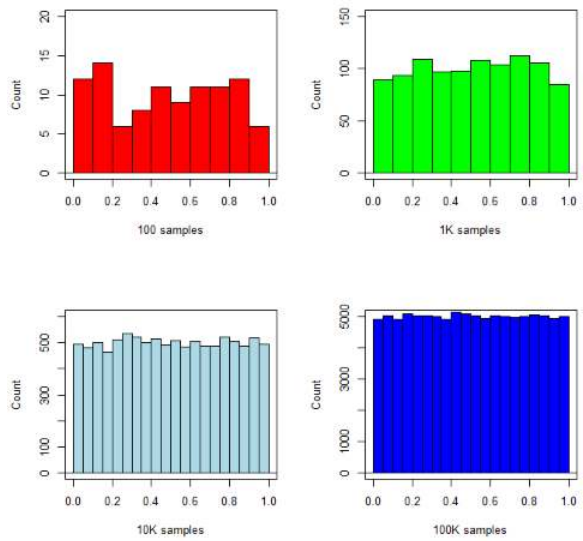


Figure 19.3.1: Histograms of runif results with 100, 1K, 10K, and 100K numbers of values to be generated.

Looks pretty uniform. A property of random numbers is that history should not influence the future, i.e., no **autocorrelation**. We can check using the `acf()` function (Fig. 19.3.2).

```
par(mfrow = c(2, 2))
acf(myUniformH, main="100")
acf(myUniform1K, main="1K")
acf(myUniform10K, main="10K")
acf(myUniform100K, main="100K"
dev.off()
```

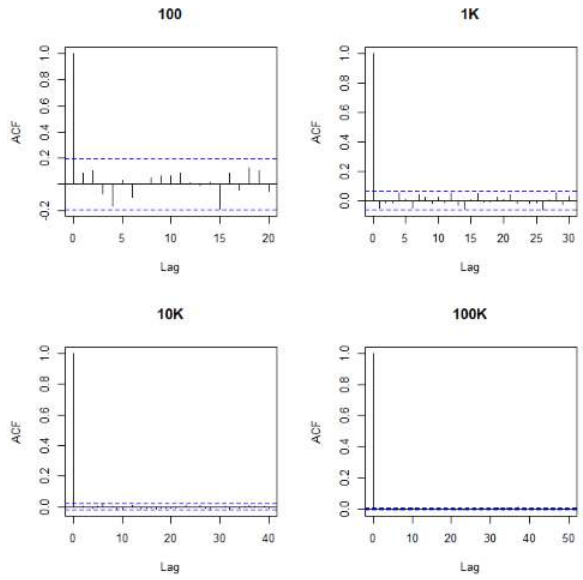


Figure 19.3.2: Autocorrelation plots of runif results with 100, 1K, 10K, and 100K numbers of values.

Correlations among points are plotted versus lag, where **lag** refers to the number of points between adjacent points, e.g., lag = 10 reflects the correlation among points 1 and 11, 2 and 12, and so forth. The band defined by two parallel blue dashed lines

Questions

1. Use `set.seed(123)` and repeat `runif(10)` twice. Confirm that the two sets are different (do not set seed) or the same when `set.seed` is used. R hint: use function `identical(x,y)` , where x and y are the two generated samples. This function tests whether the values and sequence of elements are the same between the two vectors.

19.4: References and suggested reading

Kulesa, A., Krzywinski, M., Blainey, P. *et al.* Sampling distributions and the bootstrap. *Nat Methods* **12**, 477–478 (2015). https://doi.org/10.1038/nmeth.3414.

Severiano, A., Carriço, J. A., Robinson, D. A., Ramirez, M., & Pinto, F. R. (2011). Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. *PLoS One*, *6*(5), e19539. https://doi.org/10.1371/journal.pone.0019539.

CHAPTER OVERVIEW

20: Additional Topics

under construction — add brief descriptions

Introduction

Biostatistics covers a wide variety of applied topics. This final chapter contains brief annotations and R script for the following additional topics:

20.1: Area under the curve

Introduction

Area under the curve, **AUC**, represents the total change in y given change in x. For example, if x is time, and y is oxygen consumption, an AUC would be appropriate to quantify the total oxygen consumption following strenuous exercise (Excess post-exercise oxygen consumption, EPOC) or following a large meal (Specific Dynamic Action, SDA).

In biostatistics, **area under the relative (receiver) operating carrier, AUROC**, shows characteristics of a diagnostic model, a graphic used to show tradeoff between sensitivity and specificity. Classifier performance. Used to find the appropriate cut-off. Plot true positive rates against false positive rates as **cumulative functions**, shows the relationship between sensitivity and specificity for every possible **cut off value**. Can then calculate AUC to get a measure of the intervention's ability to discriminate between true and false positive rates.

edit

Related, **area under precision-recall curve, AUPRC**,

estimate area (1) trapezoid method, (2) average precision score

Area under the curve

Download and install R package `MESS` ; requires `geepack` , `geeM` , and `Matrix` packages

R code

```
x <- seq(1:10)
y <- c(1,4,5,2,11,22,9,7,5,1)
#length(x)==length(y)
#smooth the data
loxy <- loess(y~x)
#Make a plot (Fig. 20.1.1)
plot(x,y, pch=19, cex=2, col="blue")
lines(predict(loxy), type="l", col="red")
```

where == is an R **comparison operator**.

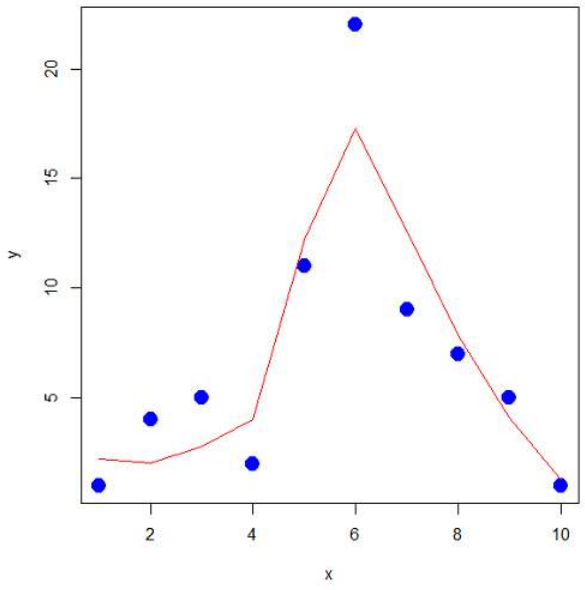


Figure 20.1.1: Area under the curve example.

```
library(MESS)
auc(x,y,from=0,rule=2)
auc(x,loxy$fitted,from=0,rule=2)
```

And R output

```
#area under curve for raw data
[1] 67
#area under curve for smoothed data
[1] 66.77616
```

Area under the receiver operating carrier curve

Download and install ROCR

R code

#modified from https://rviews.rstudio.com/2019/03/01/some-r-packages-for-roc-curves/

```
library(ROCR)
data(ROCR.simple)
df <- data.frame(ROCR.simple)
pred <- prediction(df$predictions, df$labels)
perf <- performance(pred,"tpr","fpr")
plot(perf,colorize=TRUE)
```

R output:

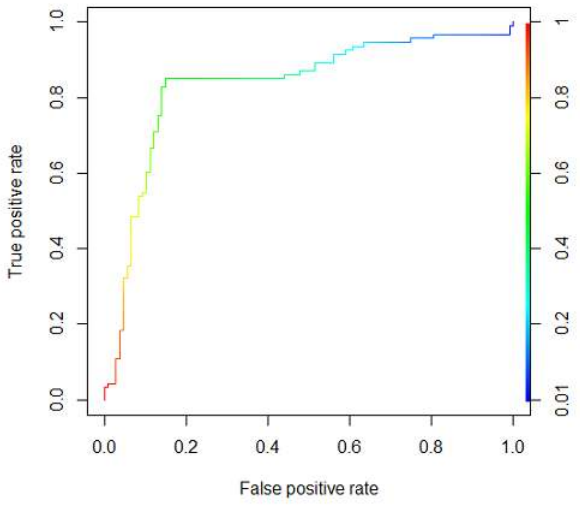


Figure 20.1.2: Example ROC curve.

The right-hand axes is color codes by AUC values: good tests AUC between 0.8 and 0.9, very good tests greater than 0.9.

Area under the precision recall curve

— under construction

Questions

[pending]

20.2: Peak detection

[under construction]

Introduction

algorithm

extract characteristics

shape

signal

noise

intensity

filtering

window length

R code

packages

peakDetection

findpeaks

Example

ccc

Questions

[pending]

References and suggested readings

Shin, H. S., Lee, C., & Lee, M. (2009). Adaptive threshold method for the peak detection of photoplethysmographic waveform. *Computers in biology and medicine*, *39*(12), 1145-1152.

20.3: Baseline correction

[under construction]

Introduction

distortion, background noise, baseline drift

baseline itself is an estimate

signal, baseline windows

regression-weighted correction

spline

R code

Package(s):

baseline

Examples

[pending]

Questions

[pending]

References

Liland, K. H. (2015). 4S Peak Filling–baseline estimation by iterative mean suppression. *MethodsX*, 2, 135-140.

Liland, K. H., & Mevik, T. A. B. H. (2011). Optimal baseline correction for multivariate calibration using open-source software. *Life Science Instruments*, (3), 7.

20.4: Conducting surveys

under construction, missing citations

Introduction

What is a survey? A **survey** is a method of collecting information from a sample of a **reference population**. Surveys are implemented in many fields including biomedical work (see Chapter 5.4). We can design a **cross-sectional** study, which gathers information at one time, or the study can be **longitudinal**, whereby information is gathered over a period of time. If the purpose of the survey is to determine **association** between two (or more) variables, then either cross-sectional or longitudinal approaches will do. However, if **cause-effect hypotheses** are the purpose, then longitudinal (e.g., **prospective cohort**) approaches would be better.

Survey basics

Steps to conduct a survey include

1. Identify and clarify the purpose of the survey. If the purpose is to find out how common something is, then this is descriptive. If we are interested in why something has occurred, then this is an analytic survey.
2. Define the reference population. It is essential that you know which group the survey applies to. For example, if one wishes to study the opinions of undergraduates at your university, then postgraduate students cannot be included in the sample as they are not part of the reference population.
3. Design sampling method and determine sample size. Sampling needs to be done to obtain unbiased sample from representative population. If the size of the population is known, then a target of 10% might be the relevant sample size, and a procedure should be taken to obtain a simple random sample. A measure of the success of a survey sample design is the size of the response rate, defined as the ratio of surveys returned divided by the number of surveys distributed.
4. What information is needed? Care needs to be taken to make sure that the questions asked actually yield the desired information. For example, if the questionnaire is long, there may be a tendency for some to skim or skip questions. If too much information is requested, this may lower the response rate.
5. How will the information be collected? Types and format of questions (closed or open ended). A phone survey? Written survey dropped off as a mailer? Interview?
6. In thinking about the data to be collected from the questions, you also need some scoring system. Will a dichotomous response (e.g., True/False, or Yes/No) be adequate, or would a Likert-like scale be more appropriate? When scaling responses, one needs to also be concerned with floor and ceiling effects.
7. Collect the data. What protocol will be employed to achieve a high response rate with unbiased responses?
8. Analyze the data. Often chi-square contingency table or the related logistic regression would be appropriate.

Bias sources in survey research

When a statistical estimator consistently under or over estimates the true value, this is called **statistical bias**, which was introduced and discussed in Chapter 5.3. The potential for bias responses to survey questions is an important constraint on the applicability of the survey. An example, it is well-known that adults tend to overestimate their height, but underestimate their weight.

In survey research, different classes of bias have been defined.

- **Information bias** occurs when trends are present in the measurement of the response, (1) recall bias and (2) observer bias.
- **Recall bias** is when a difference occurs because some people are much more likely to remember and event than others.
- **Observer bias** can be as a result of differences between different observers. If different persons conduct interviews, then it is important that all observers use a standardized method of collecting data.
- A reality of survey is that not all targets of the survey will answer the questions. This may result in bias. **Non-response bias** is the situation when those who respond to a questionnaire, the **responders**, differ in some way from those who don't, the **non-responders**.
- **Selection bias** results when the sample group you have chosen is not representative of the population you want to generalize your results to. Random sampling can help to minimize this from happening in your survey, but a stratified sampling approach is needed to avoid missing representation, e.g., economic groups, ethnicity.

How to ask questions?

The goal is to maximize the number of people who respond to the survey while maintaining accuracy and relevance of the responses. This is accomplished by asking the right questions in the right manner, but also by how the questionnaire is presented

and administered. To get accurate answers, one should include additional questions to check the consistency of the responses provided by a person. For example, if the study is about smoking, you could ask either of two questions (or both…?):

Question 1. Do you smoke tobacco cigarettes? Yes / No

Question 2. How many cigarettes did you smoke yesterday?
0 …. 1 – 10 …. 11 – 20 …. 21 +

Note that these questions are CLOSED — the responder must answer using the answers provided rather than making something up. This has the advantage of restricting the possible answers and allows you to test specific hypotheses. An OPEN question might be

Question 3. Do you smoke a lot of cigarettes in a day?

As you can imagine, you would expect to get a variety of interpretations of this question, which limits your ability to analyze test the hypothesis. It would be a poor question to use.

Closed or forced format questions can take on a variety of styles.

Question 4. What is your favorite soft drink? (select one answer only)
__ water
__ cola
__ ice tea
__ fruit juice
__ no preference

Note that Question 4 gives the responder a choice among categories. For another example,

Question 5. Biostatistics is my favorite subject
___Strongly disagree ___Disagree ___Undecided ___Agree ___Strongly Agree

And still another example might employ ranking, requesting the responder to rank from 1 to 8 their favorite subjects from a list of academic topics.

All of these closed format responses can be easily converted for analyses by contingency table or other nonparametric statistics.

Some other general tips in writing a survey.

- Keep sentences simple and short.
- Ask for only one piece of information at a time.
- Ask precise questions.
- Start the survey with the question(s) most relevant to the subject of the study.
- Avoid asking personal questions at the start.
- Write some questions then conduct a pilot study to test for the efficacy of the survey

Suggested readings

Statistics Canada: a site with lots of material about survey methods.

Wikipedia: Statistical survey

20.5: Time series

Introduction

Time series refer to any measure recorded over time. **Stationary time series** do not have trends or seasonality, just random (white) noise; **differencing time series** do have trends and or seasonality. Stationary time series will not have predictable patterns over the long term.

This page is under construction. Examples and questions are in place, but not much else; here's a resource on time series:

http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

R code

To conduct time series analysis use built in functions like `ts()` and `decompose()`. `HoltWinters()` also useful, now part of stats. Lots of specialized time series packages with advanced features, including `forecast`, `timeSeries` (Financial time series), `season` (Seasonal analysis of health data), and many others.

Note:

Note 1: Caution — newer versions of R have `HoltWinters()` and related functions included with base package `stats`.

Note 2: Rcmdr package for time series was RcmdrPlugin.epack, no longer available as of 2018.

For up-to-date listing of time series packages, see https://cran.r-project.org/web/views/TimeSeries.html

Time series data sets included in R and Rcmdr

R Code:

```
data(co2, package="datasets")
co2 <- as.data.frame(co2)
```

```
#convert to time series data type with ts()
tCO2 <- ts(co2,frequency=12,start=c(1959),end=c(1997))
plot.ts(tCO2)
```

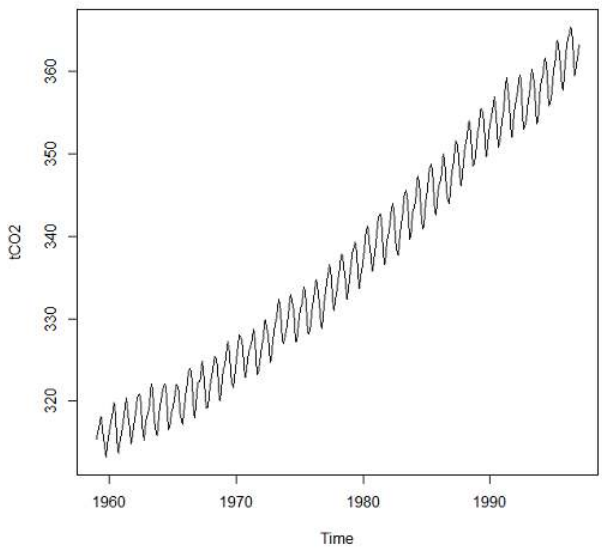


Figure 20.5.1: CO_2 data set from package datasets, comes with Rcmdr installation.

Other datasets included with R

carData::Arrests

carData::Bfox

carData::CanPop

Note: Dr D needs to complete this list

Example

Get up-to-date CO_2 data from NOAA as text file. Download to your computer, load and clean in your favorite spreadsheet app. Months came as numbers 1,2,3, etc., I changed to text, Jan, Feb, Mar, etc. I grabbed three columns: year, month, ppm for import to R.

```
head(maunaLoa)
```

R output:

```
> head(maunaLoa)
  year month  ppm
1 1958 Mar 315.70
2 1958 Apr 317.45
3 1958 May 317.51
4 1958 Jun 317.24
5 1958 Jul 315.86
6 1958 Aug 314.93
```

However, it turns out the time series functions are easiest to work if only the ppm data are included.

```
tCO2 <- ts(maunaLoa[,"ppm"],frequency=12,start=c(1958,3),end=c(2020,10))
head(tCO2)
```

R output:

```
> head(tCO2)
        Mar    Apr    May    Jun    Jul    Aug
1958 315.70 317.45 317.51 317.24 315.86 314.93
```

Get our plot (Figure 20.5.2).

```
plot(tCO2)
```

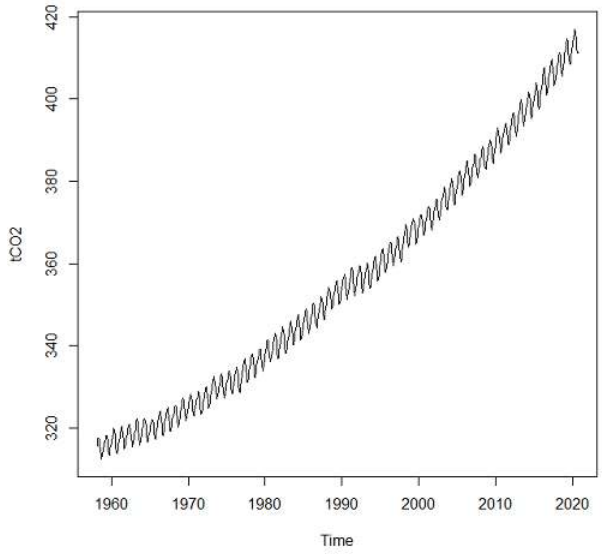


Figure 20.5.2: CO_2 ppm monthly average data from NOAA, last data October 2020.

Seasonal time series come with a trend component, a seasonal component, and a random component.

R code:

```
dectCO2 <- decompose(tCO2)
head(dectCO2)
plot(dectCO2)
```

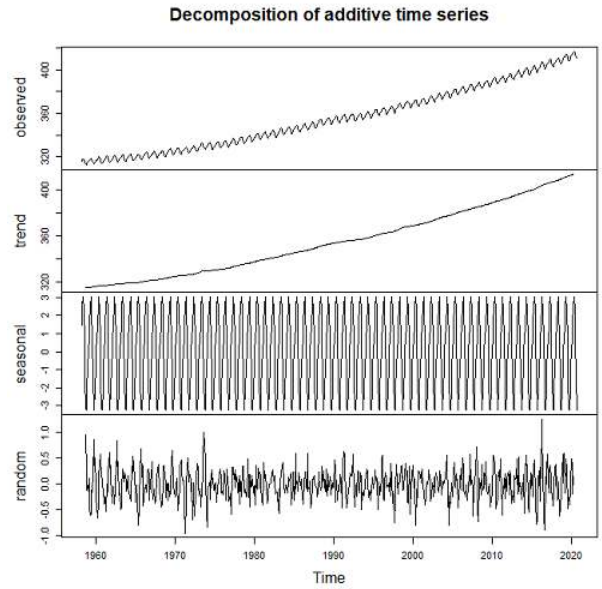


Figure 20.5.3: Observed (panel, top), trends over time (panel, second from top), seasonal changes (panel, second from bottom), and random error (panel, bottom).

Forecasting

Excellent resource at https://otexts.com/fpp2/

Exponential smoothing, weighted averages of past observations, weighted so that more recent observations are more influential.

Holt-Winters method extracts seasonal component (additive or multiplicative).

```
#set start value to value of first observation
tCO2cast <- HoltWinters(tCO2, l.start=315.42)
```

```
#Predict for next ten years. Because frequency in ts() was monthly, ten years is h=12(
forecastCO2 <- forecast(tCO2cast, h=120)
plot(forecastCO2, fcol="red")
```

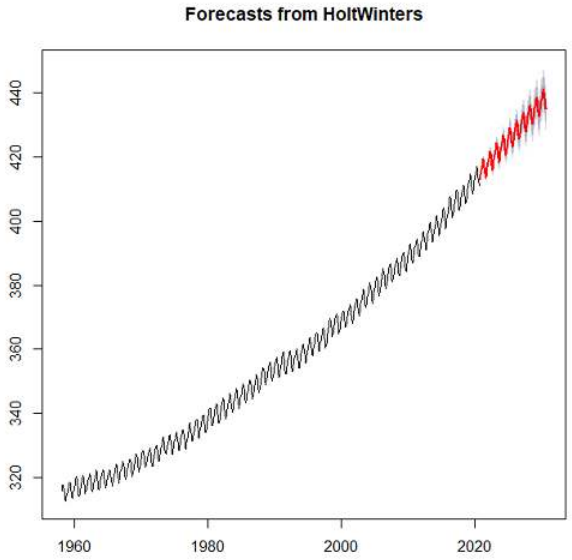


Figure 20.5.4: Data in black, predicted values in red (additive) shaded by confidence interval.

ARIMA models

DrD needs to complete

Questions

1. If a time series data set obtains observations collected at yearly intervals, what value should you enter in `ts()` function for frequency?
2. For the CO_2 dataset included in Rcmdr (co2, datasets), obtain forecast for year 2020 and compare against actual 2020 data (see Figure 20.5.2).
3. Positive clinical samples between September 2015 and November 2020 for flu virus in the USA are provided in the data set below (scroll or click here). The frequency of observations was weekly. Apply `decompose()` and obtain the seasonal and trend components of the data set. Which month does the peak positive sample occur?
4. Total pounds of fish (variable = Pounds) and pounds of Akule and Opelu (variable = Akule.Opelu) caught by commercial industry in Hawaii, from 2000 to 2018 are provided in the data set below (scroll or click here). Apply `decompose()` and obtain the seasonal and trend components of the data set for Total pounds and again for Akule (*Selar crumenophthalmus*) and Opelu (*Decapterus macarellus*). Is there evidence for trends, and if so, describe the trend. Is there evidence of seasonality? If so, which month did peak fishing occur?

Flu data set this page

Flu, extracted 28 Nov 2020 from https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Year	Date	Week	Positive
2015	09/28/15	40	1.05578
2015	10/05/15	41	1.29662

2015	10/12/15	42	1.10855
2015	10/19/15	43	1.10807
2015	10/26/15	44	1.12344
2015	11/02/15	45	1.38224
2015	11/09/15	46	1.19344
2015	11/16/15	47	1.38506
2015	11/23/15	48	1.39498
2015	11/30/15	49	1.47544
2015	12/07/15	50	2.51181
2015	12/14/15	51	2.287
2015	12/21/15	52	2.45958
2016	01/04/16	1	2.93137
2016	01/11/16	2	4.25384
2016	01/18/16	3	5.48463
2016	01/25/16	4	6.95974
2016	02/01/16	5	9.69858
2016	02/08/16	6	12.5491
2016	02/15/16	7	15.5359
2016	02/22/16	8	18.3621
2016	02/29/16	9	21.1098
2016	03/07/16	10	23.6454
2016	03/14/16	11	19.972
2016	03/21/16	12	18.4709
2016	03/28/16	13	16.2265
2016	04/04/16	14	14.0164
2016	04/11/16	15	13.2362
2016	04/18/16	16	12.3464
2016	04/25/16	17	10.2615
2016	05/02/16	18	8.12094
2016	05/09/16	19	6.68559
2016	05/16/16	20	5.81108
2016	05/23/16	21	4.71918
2016	05/30/16	22	3.0595
2016	06/06/16	23	3.02006
2016	06/13/16	24	1.82927

2016	06/20/16	25	1.71228
2016	06/27/16	26	1.22261
2016	07/04/16	27	0.903312
2016	07/11/16	28	0.869153
2016	07/18/16	29	0.849185
2016	07/25/16	30	0.781793
2016	08/01/16	31	0.933921
2016	08/08/16	32	0.900745
2016	08/15/16	33	0.803482
2016	08/22/16	34	1.40485
2016	08/29/16	35	1.67771
2016	09/05/16	36	1.46146
2016	09/12/16	37	1.51255
2016	09/19/16	38	1.74135
2016	09/26/16	39	1.78369
2016	10/03/16	40	1.56951
2016	10/10/16	41	1.35914
2016	10/17/16	42	1.40304
2016	10/24/16	43	1.50862
2016	10/31/16	44	1.91569
2016	11/07/16	45	2.20089
2016	11/14/16	46	2.57608
2016	11/21/16	47	3.34773
2016	11/28/16	48	3.3191
2016	12/05/16	49	4.25987
2016	12/12/16	50	6.68342
2016	12/19/16	51	10.7819
2016	12/26/16	52	13.9993
2017	01/02/17	1	13.3436
2017	01/09/17	2	15.373
2017	01/16/17	3	18.2865
2017	01/23/17	4	18.5299
2017	01/30/17	5	21.4215
2017	02/06/17	6	24.1525
2017	02/13/17	7	24.5117

2017	02/20/17	8	24.7251
2017	02/27/17	9	19.772
2017	03/06/17	10	19.2714
2017	03/13/17	11	19.0338
2017	03/20/17	12	19.7113
2017	03/27/17	13	18.4816
2017	04/03/17	14	15.4251
2017	04/10/17	15	12.7395
2017	04/17/17	16	9.69626
2017	04/24/17	17	6.76776
2017	05/01/17	18	5.91752
2017	05/08/17	19	5.33264
2017	05/15/17	20	4.86286
2017	05/22/17	21	4.35223
2017	05/29/17	22	4.16524
2017	06/05/17	23	3.38586
2017	06/12/17	24	3.06229
2017	06/19/17	25	2.64932
2017	06/26/17	26	2.53401
2017	07/03/17	27	2.17791
2017	07/10/17	28	2.16392
2017	07/17/17	29	1.83895
2017	07/24/17	30	1.80607
2017	07/31/17	31	1.94796
2017	08/07/17	32	1.90048
2017	08/14/17	33	1.34281
2017	08/21/17	34	1.43382
2017	08/28/17	35	1.93535
2017	09/04/17	36	1.88806
2017	09/11/17	37	1.89622
2017	09/18/17	38	1.66942
2017	09/25/17	39	1.70313
2017	10/02/17	40	2.20191
2017	10/09/17	41	2.08975
2017	10/16/17	42	2.17647

2017	10/23/17	43	2.58279
2017	10/30/17	44	3.60729
2017	11/06/17	45	4.24472
2017	11/13/17	46	5.29966
2017	11/20/17	47	7.0877
2017	11/27/17	48	7.30533
2017	12/04/17	49	10.7453
2017	12/11/17	50	15.3549
2017	12/18/17	51	22.777
2017	12/25/17	52	25.3864
2018	01/01/18	1	25.3653
2018	01/08/18	2	26.9421
2018	01/15/18	3	27.034
2018	01/22/18	4	27.3698
2018	01/29/18	5	27.0643
2018	02/05/18	6	26.9981
2018	02/12/18	7	26.1174
2018	02/19/18	8	22.6155
2018	02/26/18	9	18.4867
2018	03/05/18	10	15.6938
2018	03/12/18	11	15.5813
2018	03/19/18	12	15.328
2018	03/26/18	13	15.1135
2018	04/02/18	14	12.6888
2018	04/09/18	15	11.2486
2018	04/16/18	16	9.39813
2018	04/23/18	17	7.99876
2018	04/30/18	18	6.25914
2018	05/07/18	19	4.39311
2018	05/14/18	20	3.16606
2018	05/21/18	21	2.39003
2018	05/28/18	22	1.52934
2018	06/04/18	23	1.57683
2018	06/11/18	24	1.29914
2018	06/18/18	25	1.02329

2018	06/25/18	26	1.11356
2018	07/02/18	27	1.00305
2018	07/09/18	28	0.916118
2018	07/16/18	29	1.0534
2018	07/23/18	30	0.995099
2018	07/30/18	31	0.953592
2018	08/06/18	32	0.95729
2018	08/13/18	33	0.764331
2018	08/20/18	34	1.33625
2018	08/27/18	35	1.50367
2018	09/03/18	36	1.74739
2018	09/10/18	37	1.68745
2018	09/17/18	38	1.69929
2018	09/24/18	39	1.49699
2018	10/01/18	40	1.74855
2018	10/08/18	41	1.6967
2018	10/15/18	42	1.99298
2018	10/22/18	43	2.05527
2018	10/29/18	44	2.17372
2018	11/05/18	45	2.7331
2018	11/12/18	46	3.15674
2018	11/19/18	47	3.92782
2018	11/26/18	48	3.91485
2018	12/03/18	49	6.23152
2018	12/10/18	50	10.3644
2018	12/17/18	51	14.2649
2018	12/24/18	52	16.352
2019	12/31/18	1	12.1387
2019	01/07/19	2	12.7217
2019	01/14/19	3	16.3174
2019	01/21/19	4	19.3918
2019	01/28/19	5	22.5493
2019	02/04/19	6	25.1342
2019	02/11/19	7	26.026
2019	02/18/19	8	26.2407

2019	02/25/19	9	26.0743
2019	03/04/19	10	25.6065
2019	03/11/19	11	26.1318
2019	03/18/19	12	22.4805
2019	03/25/19	13	19.3035
2019	04/01/19	14	14.9422
2019	04/08/19	15	11.9093
2019	04/15/19	16	8.61102
2019	04/22/19	17	5.84355
2019	04/29/19	18	4.81976
2019	05/06/19	19	3.83986
2019	05/13/19	20	3.54159
2019	05/20/19	21	3.41968
2019	05/27/19	22	3.0826
2019	06/03/19	23	2.78989
2019	06/10/19	24	2.31579
2019	06/17/19	25	1.90194
2019	06/24/19	26	2.0806
2019	07/01/19	27	2.42883
2019	07/08/19	28	2.01653
2019	07/15/19	29	2.21849
2019	07/22/19	30	2.37706
2019	07/29/19	31	2.39817
2019	08/05/19	32	2.05446
2019	08/12/19	33	2.08183
2019	08/19/19	34	2.36167
2019	08/26/19	35	3.45517
2019	09/02/19	36	3.09749
2019	09/09/19	37	2.48391
2019	09/16/19	38	2.75656
2019	09/23/19	39	2.74367
2019	09/30/19	40	1.30976
2019	10/07/19	41	1.47877
2019	10/14/19	42	1.55203
2019	10/21/19	43	2.25335

2019	10/28/19	44	3.05701
2019	11/04/19	45	5.16261
2019	11/11/19	46	6.75594
2019	11/18/19	47	9.54599
2019	11/25/19	48	10.9385
2019	12/02/19	49	11.6554
2019	12/09/19	50	16.1542
2019	12/16/19	51	22.533
2019	12/23/19	52	26.9336
2020	12/30/19	1	23.4883
2020	01/06/20	2	23.1187
2020	01/13/20	3	26.0826
2020	01/20/20	4	28.2813
2020	01/27/20	5	30.1465
2020	02/03/20	6	30.2596
2020	02/10/20	7	29.675
2020	02/17/20	8	28.3215
2020	02/24/20	9	25.7517
2020	03/02/20	10	22.4914
2020	03/09/20	11	15.8125
2020	03/16/20	12	7.50171
2020	03/23/20	13	2.32158
2020	03/30/20	14	1.0312
2020	04/06/20	15	0.61823
2020	04/13/20	16	0.623139
2020	04/20/20	17	0.218375
2020	04/27/20	18	0.262953
2020	05/04/20	19	0.326173
2020	05/11/20	20	0.305966
2020	05/18/20	21	0.212681
2020	05/25/20	22	0.16518
2020	06/01/20	23	0.339751
2020	06/08/20	24	0.279818
2020	06/15/20	25	0.38117
2020	06/22/20	26	0.282336

2020	06/29/20	27	0.210322
2020	07/06/20	28	0.176197
2020	07/13/20	29	0.37594
2020	07/20/20	30	0.150451
2020	07/27/20	31	0.132626
2020	08/03/20	32	0.176141
2020	08/10/20	33	0.132385
2020	08/17/20	34	0.226904
2020	08/24/20	35	0.314861
2020	08/31/20	36	0.201675
2020	09/07/20	37	0.186246
2020	09/14/20	38	0.39985
2020	09/21/20	39	0.224669
2020	09/28/20	40	0.330089
2020	10/05/20	41	0.400802
2020	10/12/20	42	0.350483
2020	10/19/20	43	0.25138
2020	10/26/20	44	0.201148
2020	11/02/20	45	0.176706
2020	11/09/20	46	0.221837

Fish data set in this page

Fish, Hawaii state DLNR, Pounds refers to total catch, Akule.Opelu refers to pounds for the two kinds of fish

Year	Month	Pounds	Akule.Opelu
1999	Jan	2064023	85331
1999	Feb	2286785	89537
1999	Mar	2083789	112897
1999	Apr	2446840	136301
1999	May	2300842	103692
1999	Jun	2340116	134432
1999	Jul	2646429	138814
1999	Aug	2254408	96569
1999	Sep	1926381	56598
1999	Oct	2233789	76834
1999	Nov	1730672	134706

1999	Dec	1762375	92255
2000	Jan	1501164	147104
2000	Feb	1993373	104165
2000	Mar	2220831	132028
2000	Apr	2398180	119224
2000	May	2557229	121268
2000	Jun	2510298	145200
2000	Jul	2270954	93883
2000	Aug	1912654	69107
2000	Sep	1365264	65007
2000	Oct	1615117	51208
2000	Nov	1388453	117493
2000	Dec	1802926	121486
2001	Jan	1481810	170702
2001	Feb	1496356	44575
2001	Mar	1579528	101764
2001	Apr	1184591	89388
2001	May	2091424	124193
2001	Jun	1966886	61122
2001	Jul	2113931	73266
2001	Aug	1926661	29386
2001	Sep	1353429	30268
2001	Oct	1338289	29577
2001	Nov	1747198	80350
2001	Dec	1458336	22817
2002	Jan	1517609	107406
2002	Feb	1729084	31030
2002	Mar	1747985	67691
2002	Apr	2109451	101043
2002	May	2069921	57251
2002	Jun	1640151	100501
2002	Jul	1979382	87584
2002	Aug	1831678	65566
2002	Sep	1734201	53162
2002	Oct	1779207	93867

2002	Nov	2191825	106167
2002	Dec	2576191	67881
2003	Jan	1910500	49420
2003	Feb	2075168	55006
2003	Mar	2245753	71616
2003	Apr	1562751	102993
2003	May	2440228	106600
2003	Jun	1842907	101715
2003	Jul	1957279	48453
2003	Aug	2143823	69130
2003	Sep	1503212	74525
2003	Oct	1611779	70949
2003	Nov	1668167	54004
2003	Dec	2312537	43054
2004	Jan	1605595	75751
2004	Feb	1705533	94864
2004	Mar	2079402	120305
2004	Apr	1883704	90950
2004	May	1830168	111599
2004	Jun	1918622	76392
2004	Jul	2029787	98937
2004	Aug	1928009	72577
2004	Sep	1620224	82650
2004	Oct	1854643	74587
2004	Nov	1981567	59753
2004	Dec	2022272	44353
2005	Jan	2088821	60972
2005	Feb	2106948	59469
2005	Mar	2386327	84551
2005	Apr	2122171	101099
2005	May	2369953	79042
2005	Jun	2342117	104814
2005	Jul	2281871	71065
2005	Aug	2124303	53383
2005	Sep	1734986	37195

2005	Oct	1920131	48632
2005	Nov	1969506	88235
2005	Dec	2323933	98768
2006	Jan	1702766	50553
2006	Feb	2060204	89037
2006	Mar	2244570	33916
2006	Apr	2068922	74430
2006	May	2164076	108689
2006	Jun	1935951	89503
2006	Jul	1968513	93758
2006	Aug	1741802	111080
2006	Sep	1508897	44537
2006	Oct	1892535	46747
2006	Nov	2208173	82938
2006	Dec	1381412	42260
2007	Jan	2211384	114496
2007	Feb	2391437	60618
2007	Mar	2724021	94251
2007	Apr	2639245	90078
2007	May	3168913	129258
2007	Jun	2706972	116628
2007	Jul	2523392	129345
2007	Aug	2272502	88997
2007	Sep	2121837	71560
2007	Oct	2472996	52915
2007	Nov	3040118	107555
2007	Dec	2934174	39239
2008	Jan	2656539	44672
2008	Feb	3101819	35213
2008	Mar	2816846	74421
2008	Apr	3064837	63355
2008	May	3560993	52287
2008	Jun	2920219	33685
2008	Jul	2516561	31288
2008	Aug	2338205	62171

2008	Sep	2314458	31311
2008	Oct	2407240	42766
2008	Nov	2060666	75102
2008	Dec	2329268	74508
2009	Jan	2198569	44459
2009	Feb	2314764	33206
2009	Mar	1846459	64879
2009	Apr	2659230	36638
2009	May	2692440	77011
2009	Jun	2387175	49217
2009	Jul	2672895	55033
2009	Aug	2174027	40398
2009	Sep	2259153	51386
2009	Oct	2386749	58095
2009	Nov	2081706	51798
2009	Dec	2702871	55148
2010	Jan	2059964	40855
2010	Feb	2632985	100598
2010	Mar	2430562	39887
2010	Apr	2652013	40528
2010	May	2460228	71483
2010	Jun	2743053	120553
2010	Jul	2278847	96315
2010	Aug	2618427	62854
2010	Sep	2483861	66613
2010	Oct	2503321	53353
2010	Nov	2370032	104360
2010	Dec	2431047	57919
2011	Jan	2527241	37755
2011	Feb	2786453	51863
2011	Mar	3789076	40188
2011	Apr	3148826	60494
2011	May	3015187	49037
2011	Jun	2718583	58380
2011	Jul	2284521	43096

2011	Aug	2475519	33612
2011	Sep	2461640	48697
2011	Oct	2420554	49929
2011	Nov	2059769	63045
2011	Dec	2882776	64430
2012	Jan	2825116	42894
2012	Feb	2653892	23528
2012	Mar	2544758	39839
2012	Apr	3050109	47250
2012	May	3264666	41357
2012	Jun	2798204	56808
2012	Jul	3331174	46853
2012	Aug	2864088	62682
2012	Sep	2219536	33641
2012	Oct	2482162	47478
2012	Nov	2545142	49232
2012	Dec	3129507	35924
2013	Jan	2902748	32373
2013	Feb	2388197	21922
2013	Mar	2831279	41718
2013	Apr	2467444	54619
2013	May	3131153	57183
2013	Jun	2819983	33484
2013	Jul	3473180	44240
2013	Aug	2586863	52288
2013	Sep	2459258	38145
2013	Oct	3228317	48533
2013	Nov	2998732	53187
2013	Dec	3023918	33381
2014	Jan	2503733	31233
2014	Feb	2615184	33134
2014	Mar	2808639	38876
2014	Apr	2857514	45819
2014	May	3363746	58283
2014	Jun	2778689	54266

2014	Jul	2828847	41221
2014	Aug	3074061	39744
2014	Sep	2703440	40668
2014	Oct	2744813	37263
2014	Nov	2541143	72020
2014	Dec	3325799	44128
2015	Jan	3130822	54942
2015	Feb	2806020	45098
2015	Mar	3560866	53378
2015	Apr	3341695	43642
2015	May	3717487	70583
2015	Jun	3678283	56578
2015	Jul	3954460	53615
2015	Aug	3016100	42015
2015	Sep	2209724	38904
2015	Oct	2795409	55583
2015	Nov	3426753	70399
2015	Dec	3357454	51095
2016	Jan	3087231	54089
2016	Feb	3374485	48683
2016	Mar	3260054	45472
2016	Apr	2930106	63926
2016	May	3383331	76757
2016	Jun	3209613	45557
2016	Jul	2765143	37198
2016	Aug	2732867	40213
2016	Sep	2180347	41660
2016	Oct	2298348	34699
2016	Nov	2545574	71924
2016	Dec	3691485	37448
2017	Jan	3383297	48974
2017	Feb	2856584	35716
2017	Mar	3413039	39789
2017	Apr	3361156	30625
2017	May	3576410	31092

2017	Jun	3348469	27734
2017	Jul	2741187	27041
2017	Aug	2675625	32476
2017	Sep	2700675	33394
2017	Oct	2779159	31373
2017	Nov	2817012	40681
2017	Dec	3726216	33955
2018	Jan	3361591	46166
2018	Feb	2625263	29890
2018	Mar	3219102	31454
2018	Apr	3593287	25954
2018	May	3798285	35908
2018	Jun	3362829	31899
2018	Jul	2735326	30968
2018	Aug	2397549	19849
2018	Sep	2323735	29324
2018	Oct	2472451	28927
2018	Nov	2687466	40497
2018	Dec	3236293	36603

20.6: Dimensional analysis

draft

Introduction

Cluster analysis or clustering is a multivariate analysis technique that includes a number of different algorithms for grouping objects in such a way that objects in the same group (called a cluster) are more similar to each other than they are to objects in other groups. A number of approaches have been taken, but loosely can be grouped into **distance clustering methods** (see Chapter 16.6 – Similarity and Distance) and **linkage clustering methods**: Distance methods involve calculating the distance (or similarity) between two points and whereas linkage methods involve calculating distances among the clusters. **Single linkage** involves calculating the distance among all pairwise comparisons between two clusters, then

Cluster analysis is common to molecular biology and phylogeny construction and more generally is an approach in use for exploratory data mining. Unsupervised machine learning (see 20.14 – Binary classification) used to classify, for example, methylation status of normal and diseased tissues from arrays (Clifford et al 2011)

Results from cluster analyses are often displayed as dendrograms. Clustering methods include a number of different algorithms hierarchical clustering: single-linkage clustering; complete linkage clustering; average linkage clustering (UPGMA) centroid based clustering: k-means clustering

R packages

factoextra

`psa` package from MorphoFun/psa/

Principal component analysis

Bumpus data from MorphoFun/psa, variable names changed.

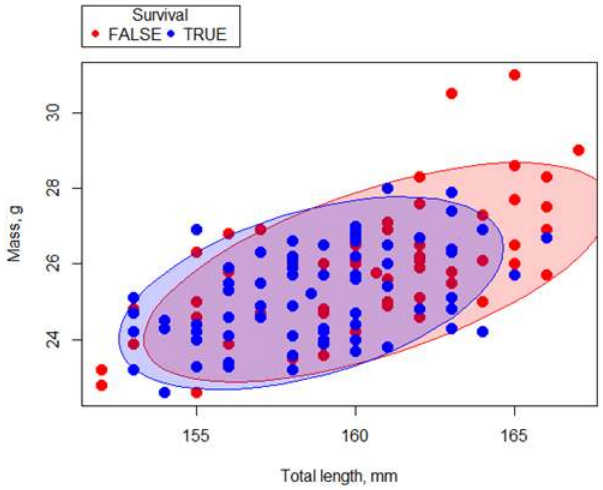


Figure 20.6.1: Scatterplot of English swallow mass (g) vs. total length (mm), by survival following winter storm.

R code for graph

```
scatterplot(Weight~Total_length | Survival, regLine=FALSE, smooth=FALSE, boxplots=FALS
ellipse=list(levels=c(.9)), by.groups=TRUE, grid=FALSE, pch=c(19,19), cex=1.5, col=c(
```

Data ellipse — 90% of the pairwise points (red, did not survive; blue, did survive), not a confidence ellipse

Bumpus measured several traits, we want to use all of the data. However, highly correlated (Fig. 20.6.2) and therefore multicollinear.

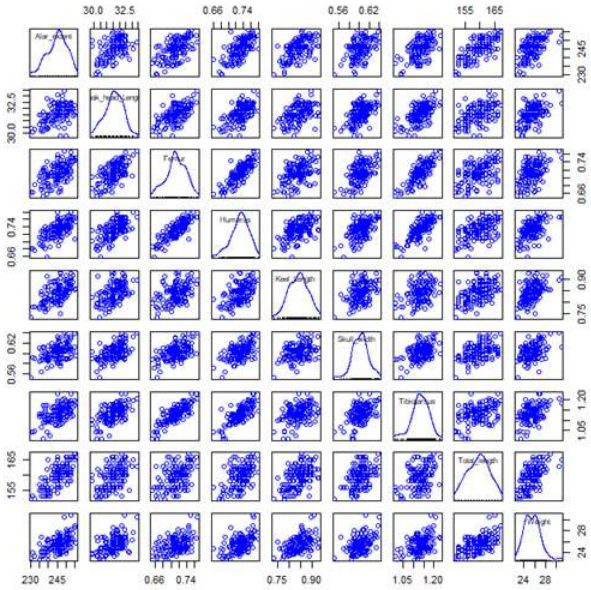


Figure 20.6.2: Scatterplot matrix of Bumpus English sparrow traits. Traits were (left-right): Alar extent (mm), length (tip of beak to tip of tail), length of head (mm), length of femur (in.), length of humerus (in.), length of sternum (in.), skull width (in.), length of tibio-taurus (in.), and weight (g).

R code for graph:

```
scatterplotMatrix(~Alar_extent+Beak_head_Length+Femur+Humerus+Keel_Length+Skull_width+
regLine=FALSE, smooth=FALSE, diagonal=list(method="density"), data=Bumpus)
```

Note:

In Chapter 4, we discussed the importance of white space and Y-scale for graphs that make comparisons. Figure 20.6.2 is a good example of where we trade-off the need for white space and concerns about telling the story — the various traits are positively correlated — against the dictum of an equal Y-scale for true comparisons.

Rcmdr: Statistics > Dimensional analysis > Principal component analysis …

```
.PC <-
princomp(~Alar_extent+Beak_head_Length+Femur+Humerus+Keel_Length+Skull_width+Tibiotar
cor=TRUE, data=Bumpus)
cat("\nComponent loadings:\n")
print(unclass(loadings(.PC)))
cat("\nComponent variances:\n")
print(.PC\$sd^2)
cat("\n")
print(summary(.PC))
screeplot(.PC)
Bumpus <<- within(Bumpus, {
PC2 <- .PC\$scores[,2]
PC1 <- .PC\$scores[,1]
})
})
```

Importance of components

```
                          Comp.1     Comp.2
Standard deviation        2.3046882  0.9988978
Proportion of Variance    0.5901764  0.1108663
Cumulative Proportion     0.5901764  0.7010427
```

K-means clustering

Number of clusters

Iterations

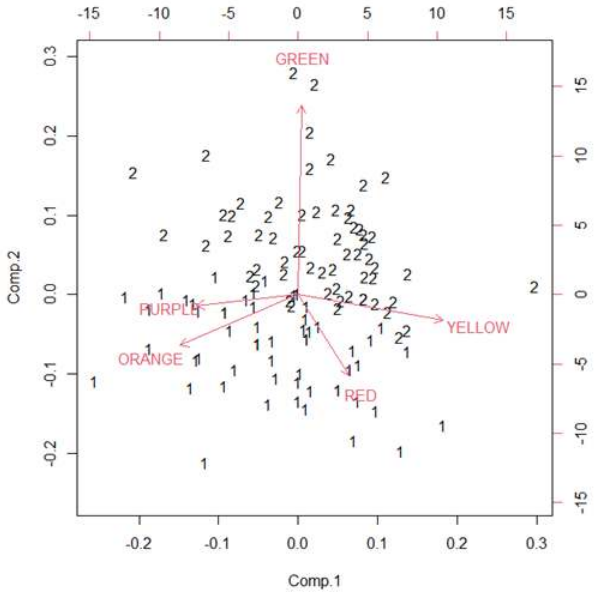


Figure 20.6.3: Bi-plot of clusters by color from Skittles mini bags.

Ward's method

Complete linkage

McQuitty's method

Centroid linkage

A common way to depict the results of a cluster analysis is to construct a dendogram.

Questions

[pending]

References and further reading

Bumpus, H. C. (1898). Eleventh lecture. The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. (A fourth contribution to the study of variation.). *Biology Lectures: Woods Hole Marine Biological Laboratory*, 209–255.

Clifford, H., Wessely, F., Pendurthi, S., & Emes, R. D. (2011). Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in genetics*, *2*, 88.

Ferreira, L., & Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, *38*(9), 1925-1949.

Fraley C, Raftery AE. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*; 97(458):611–31.

20.7: Estimating population size

[under construction]

Approaches to finding out how many individuals are present in a particular geographic area.

Census methods

If the population is closed, and all individual can be discovered, then counting every individual is the best way to estimate the population size.

Simple random and systematic sampling

Random sampling would be to divide an area into a grid then randomly select grids to be counted. Systematic sampling would be to identify areas ahead of time which are likely to have the individuals, then proceed to count individuals in all areas where the individuals are likely to be.

Capture-recapture methods

$$\frac{m}{n} = \frac{M}{N}$$

then solve for $N = n \cdot \frac{M}{m}$

This is called the Lincoln Index, where N is the estimated population size, M is the number of individuals caught the first time (and all marked, then released), n is the number of individuals captured a second time, of which m were marked. Assumptions of this method include:

1. closed population (i.e., no loss or gain of individuals during the capture intervals);
2. every individual in the population has an equal chance of being caught;
3. marks are always recognizable.

Removal methods

Using intensive methods (e.g., netting), capture animals, prevent immigration into the area. Assumption is that the captures per unit time yield decreasing numbers of caught individuals. Then, change in population size may be estimated by

$$\frac{dN}{dt} = -\alpha N$$

where α is the removal rate. The solution to this equation is the differential \(N = N_{0} e^{\alpha t\)

where e is the natural logarithm, N_0 is the initial population size, and t is time intervals. If A is the number of individuals captured at time t_i, then a plot of A on the Y-axis versus t_i describes this differential. α could be estimated by getting the slope of the non-linear regression (N_0 would be the intercept).

As an approximation, you could take the based on the analysis of 2 first time intervals only. For example, if captures in the first 2 time intervals were 23 and 14 fish, then

$$N_0 = \frac{232}{23 - 15} = 66.125$$

$$\alpha = \frac{23 - 15}{23} = 0.3478$$

Capture effort

x

Additional reading

http://www.sbs.utexas.edu/jcabbott/courses/bio208web/labs/populations/populations.htm

20.8: Diversity indexes

[under construction]

A diversity index is a measure of how many different kinds (e.g., species) are present in a dataset. These indexes are more than a count of the different types; they also account for how common (or rare) a kinds is. Diversity indexes are examples of use of multivariate statistics: two or more predictor variables and two or more response variables.

There are many varieties of diversity indexes, but two are well known.

H, Shannon's Diversity Index

Shannon's Index accounts for abundance and evenness of all species present in an area. Evenness refers to how close the numbers of each species are in an area.

Simpson

Statistical significance

Comparing indices

Software:

Download and install the BiodiversityR package, a GUI for biodiversity, suitability and community ecology analysis. Utilizes vegan package.

20.9: Survival analysis

[Page rough draft]

Introduction

As the name suggests, survival analysis is a branch of statistics used to account for death of organisms, or more generally, failure in a system. In general, this kind of analysis models time to event, where event would be death or failure. The basics of the method is defined by the **survival function**

$$S_t = Pr(T < t)$$

where t is time, T is a variable that represents time of death or other end point, and Pr is probability of an event occurring later than at time t.

Excellent resources available, series of articles in volume 89 of *British Journal of Cancer*: Clark et al (2003a), Bradburn et al (2003a), Bradburn et al (2003b), and Clark et al (2003b).

Hazard function

Defined as the event rate at time t based on survival for time times equal to or greater than t.

Censoring

Censoring is a a missing data problem typical of survival analysis. Distinguish right-censored and left-censored.

Kaplan-Meier plot

Kaplan-Meier (KM) estimator of survival function. Other survival function estimators **Fleming-Harrington**. The KM estimator, $\hat{S}_t$ is

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of events that occurred at time t_i, n_i is the number of individuals known to have survived or not been censored. Because it's an estimate, a statistic, we need an estimate of the **error variance**. Several options, the default in R is the **Greenwood** estimator.

$$var\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i\left(n_i - d_i\right)}$$

The KM plot, censoring times noted with plus.

R code

Download and install the RcmdrPlugin.survival package.

Example

```
data(heart, package="survival")
attach(heart)
#Get help with the data set
help("heart", package="survival")
```

```
head(heart)
```

```
 start stop event        age      year surgery transplant id
1     0   50     1 -17.155373 0.1232033       0          0  1
2     0    6     1   3.835729 0.2546201       0          0  2
```

```
3    0    1     0   6.297057 0.2655715       0          0  3
4    1   16     1   6.297057 0.2655715       0          1  3
5    0   36     0  -7.737166 0.4900753       0          0  4
6   36   39     1  -7.737166 0.4900753       0          1  4
```

Run basic survival analysis. After installing the `RcmdrPlugin.survival` , from Rcmdr select estimate survival function.

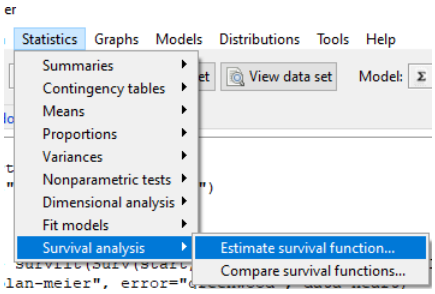


Figure 20.9.1: Screenshot of menu call for survival analysis in Rcmdr.

Get survival estimator and KM plot (Figure 20.9.2)

R output:

```
.Survfit <- survfit(Surv(start, event) ~ 1, conf.type="log", conf.int=0.95,
Rcmdr+ type="kaplan-meier", error="greenwood", data=heart)

 .Survfit
Call: survfit(formula = Surv(start, event) ~ 1, data = heart, error = "greenwood",
conf.type = "log", conf.int = 0.95, type = "kaplan-meier")

n events median 0.95LCL 0.95UCL
172 75 26 17 37

plot(.Survfit, mark.time=TRUE)

quantile(.Survfit, quantiles=c(.25,.5,.75))
$quantile
25 50 75
3 26 67

$lower
25 50 75
1 17 46

$upper
25 50 75
12 37 NA

#by default, Rcmdr removes the object
remove(.Survfit)
```

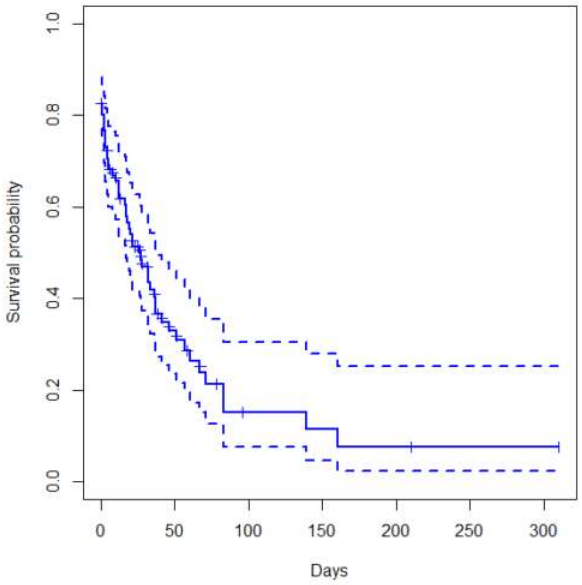


Figure 20.9.2: Kaplan-Meier plot of heart data. Dashed lines are upper and lower confidence intervals about the survival function.

Note:

I modified the `plot()` code with these additions

```
ylim = c(0,1), ylab="Survival probability", xlab="Days", lwd=2, col="blue"
```

the data set includes age and whether or not subjects had heart surgery before transplant. Compare.

Variable surgery is recorded 0,1, so need to create a factor

```
fSurgery <- as.factor(surgery)
```

Now, to compare

Rcmdr: Statistics → Survival analysis → Compare survival functions…

R output

```
Rcmdr> survdiff(Surv(start,event) ~ fSurgery, rho=0, data=heart)
Call:
survdiff(formula = Surv(start, event) ~ fSurgery, data = heart,
rho = 0)

              N Observed Expected (O-E)^2/E (O-E)^2/V
fSurgery=No 143       66     58.7     0.902      4.56
fSurgery=Yes 29        9     16.3     3.255      4.56

Chisq= 4.6 on 1 degrees of freedom, p= 0.03
```

Get the KM estimator and make a KM plot

```
mySurvfit <- survfit(Surv(start, event) ~ surgery, conf.type="log", conf.int=0.95,
type="kaplan-meier", error="greenwood", data=heart)
```

```
plot(mySurvfit, mark.time=TRUE, ylim=c(0,1),lwd=2, col=c("blue","red"), xlab="Number
legend("topright", legend = paste(c("Surgery - No", "Surgery - Yes")), col = c("blue"
```

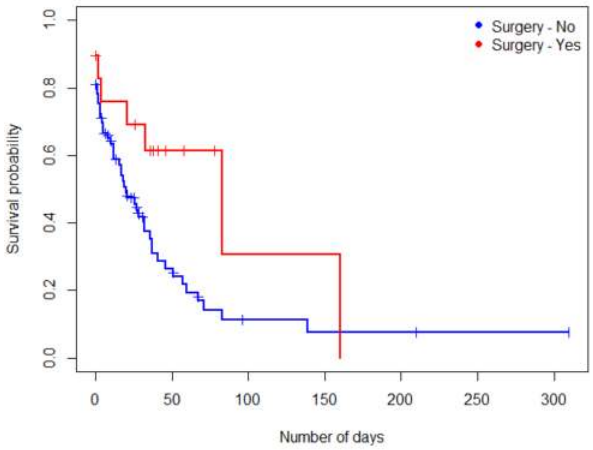


Figure 20.9.3: Kaplan-Meier plot of heart patient survival functions with and without surgery.

The comparison plot can be made in `Rcmdr` by selecting our Surgery factor in **Strata** setting (Fig. 20.9.4). Recall that strata refers to **subgroups** of a population.

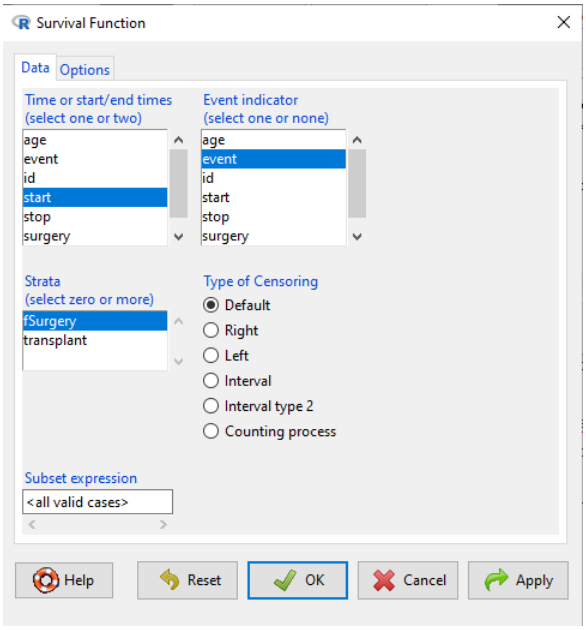


Figure 20.9.4: Screenshot of Survival estimator menu in Rcmdr.

Questions

[pending]

20.10: Growth equations and dose response calculations

Introduction

In biology, growth may refer to increase in cell number, change in size of an individual across development, or increase of number of individuals in a population over time. Nonlinear, time-series, several models proposed to fit growth data, including the Gompertz, logistic, and the von Bertalanffy. These models fit many S-shaped growth curves. These models are special cases of generalized linear models, also called Richard curves.

Growth example

This page describes how to use R to analyze growth curve data sets.

Hours	Abs
0.000	0.002207
0.274	0.010443
0.384	0.033688
0.658	0.063257
0.986	0.111848
1.260	0.249240
1.479	0.416236
1.699	0.515578
1.973	0.572632
2.137	0.589528
2.466	0.619091
2.795	0.608486
3.123	0.621136
3.671	0.616850
4.110	0.614689
4.548	0.614643
5.151	0.612465
5.534	0.606082
5.863	0.603933
6.521	0.595407
7.068	0.589006
7.671	0.578372
8.164	0.567749
8.877	0.559217
9.644	0.546451
10.466	0.537907
11.233	0.537826

Hours	Abs
11.890	0.529300
12.493	0.516551
13.205	0.505905
14.082	0.491013

Key to parameter estimates: y0 is the lag, mumax is the growth rate, and K is the asymptotic stationary growth phase. The spline function does not return an estimate for K.

R code

```
require(growthrates)
#Enter the data. Replace these example values with your own
#time variable (Hours)
Hours <- c(0.000, 0.274, 0.384, 0.658, 0.986, 1.260, 1.479, 1.699, 1.973, 2.137, 2.46
5.151, 5.534, 5.863, 6.521, 7.068, 7.671, 8.164, 8.877, 9.644, 10.466, 11.233, 11.890
#absorbance or concentration variable (Abs)
Abs <- c(0.002207, 0.010443, 0.033688, 0.063257, 0.111848, 0.249240, 0.416236, 0.5155
0.608486, 0.621136, 0.616850, 0.614689, 0.614643, 0.612465, 0.606082, 0.603933, 0.595
0.559217, 0.546451, 0.537907, 0.537826, 0.529300, 0.516551, 0.505905, 0.491013)
```

```
#Make a dataframe and check the data; If error, then check that variables have equal
Yeast <- data.frame(Hours,Abs); Yeast
```

```
#Obtain growth parameters from fit of a parametric growth model
#First, try some reasonable starting values
p <- c(y0 = 0.001, mumax = 0.5, K = 0.6)
model.par <- fit_growthmodel(FUN = grow_logistic, p = p, Hours, Abs, method=c("L-BFGS
summary(model.par)
coef(model.par)
```

```
#Obtain growth parameters from fit of a nonparametric smoothing spline
model.npar <- fit_spline(Hours,Abs)
summary(model.npar)
coef(spline.md)
```

```
#Make plots
par(mfrow = c(2, 1))
plot(Yeast, ylim=c(0,1), cex=1.5,pch=16, main="Parametric Nonlinear Growth Model", xl
lines(model.par, col="blue", lwd=2)
plot(model.npar, ylim=c(0,1), lwd=2, main="Nonparametric Spline Fit", xlab="Hours", y
```

Results from example code

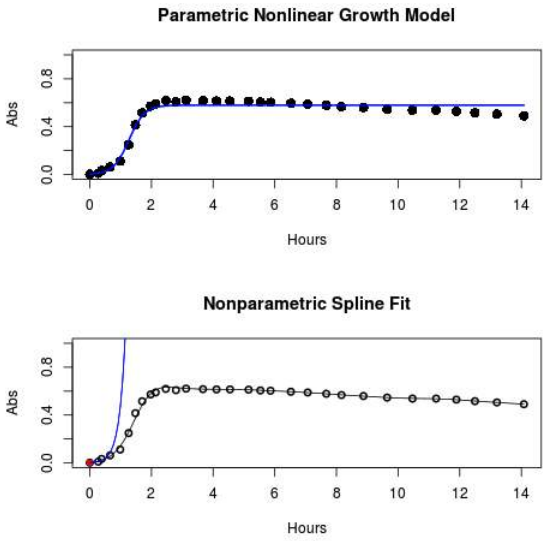


Figure 20.10.1: Top: Parametric Nonlinear Growth Model; Bottom: Nonparametric Spline Fit.

LD_{50}

In toxicology, the dose of a pathogen, radiation, or toxin required to kill half the members of a tested population of animals or cells is called the lethal dose, 50%, or LD_{50}. This measure is also known as the lethal concentration, LC_{50}, or properly after a specified test duration, the LCt_{50} indicating the lethal concentration and time of exposure. LD_{50} figures are frequently used as a general indicator of a substance's acute toxicity. A lower LD_{50} is indicative of increased toxicity.

The point at which 50% response of studied organisms to range of doses of a substance (e.g., agonist, antagonist, inhibitor, etc.) to any response, from change in behavior or life history characteristics up to and including death can be described by the methods described in this chapter. The procedures outlined below assume that there is but one inflection point, i.e., an "s-shaped" curve, either up or down; if there are more than one inflection points, then the logistic equations described will not fit the data well and other choices need to be made (see Di Veroli et al 2015). We will use the `drc` package (Ritz et al 2015).

Example

First we'll work through use of R. We'll follow up with how to use Solver in Microsoft Excel.

After starting R, load the `drc` library.

```
library(drc)
```

Consider some hypothetical 24-hour survival data for yeast exposed to salt solutions. Let `resp` equal the variable for frequency of survival (e.g., calculated from OD_{660} readings) and `NaCl` equal the millimolar (mm) salt concentrations or doses.

At the R prompt type:

```
resp <- c(1,1,1,.9,.7,.3,.4,.2,0,0,0)
NaCl=seq(0,1000,100)
#Confirm sequence was correctly created; alternatively, enter the values.
NaCl
[1] 0 100 200 300 400 500 600 700 800 900 1000
#Make a plot
plot(NaCl,resp,pch=19,cex=1.2,col="blue",xlab="NaCl [mm]",ylab="Survival frequency")
```

And here is the plot (Fig. 20.10.2).

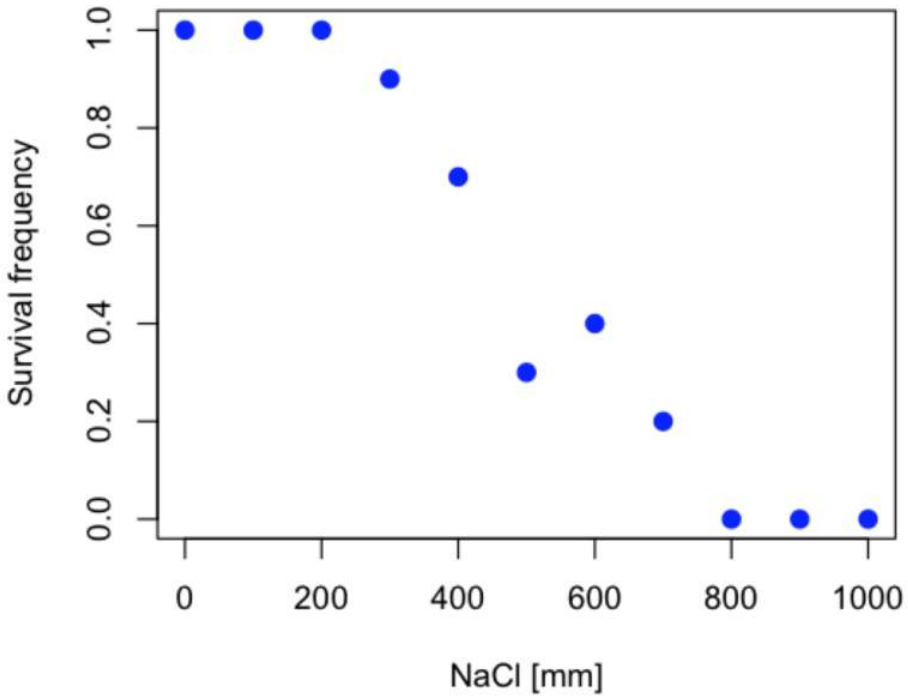


Figure 20.10.2: Hypothetical data set, survival of yeast in different salt concentrations.

Note the sigmoidal "S" shape — we'll need an logistic equation to describe the relationship between survival of yeast and NaCl doses.

The equation for the **four-parameter logistic curve**, also called the **Hill-Slope model**, is

$$f(b,c,d,e) = c + \frac{d-c}{1+exp^{b(\log(x)-\log(e))}}$$

where c is the parameter for the lower limit of the response, d is the parameter for the upper limit of the response, e is the relative EC_{50}, or the dose fitted halfway between the limits c and d, and b is the relative slope around the EC_{50}. The slope, b, is also known as the Hill slope. Because this experiment included a dose of zero, a three-parameter logistic curve would be appropriate. The equation simplifies to

$$f(b,d,e) = \frac{d}{1+exp^{b(\log(x)-\log(e))}}$$

EC_{50} from 4 parameter model

Let's first make a data frame

```
dose <- data.frame(NaCl, resp)
```

Then call up a function, `drm`, from the `drc` library and specify the model as the four parameter logistic equation, specified as `LL.4()`. We follow with a call to the summary command to retrieve output from the `drm` function. Note that the four-parameter logistic equation

```
model.dose1 = drm(dose,fct=LL.4())
summary(model.dose1)
```

And here is the R output.

```
Model fitted: Log-logistic (ED50 as parameter) (4 parms)
Parameter estimates:
                   Estimate  Std. Error    t-value  p-value
b:(Intercept)      3.753415    1.074050   3.494636   0.0101
c:(Intercept)     -0.084487    0.127962  -0.660251   0.5302
d:(Intercept)      1.017592    0.052460  19.397441   0.0000
e:(Intercept)    492.645128   47.679765  10.332373   0.0000

Residual standard error:
0.0845254 (7 degrees of freedom)
```

The EC_{50}, or technically the LD_{50} because the data were for survival, is the value of e: 492.65 mM NaCl.

You should always plot the predicted line from your model against the real data and inspect the fit.

At the R prompt type

```
plot(model.dose1, log="",pch=19,cex=1.2,col="blue",xlab="NaCl [mm]",ylab="Survival fr
```

As long as the plot you made in earlier steps is still available, R will add the line specified in the lines command. Here is the plot with the predicted logistic line displayed (Fig. 20.10.3).

Figure 20.10.3: Logistic curve added to Figure 20.10.2 plot.

While there are additional steps we can take to decide is the fit of the logistic curve was good to the data, visual inspection suggests that indeed the curve fits the data reasonably well.

More work to do

Because the EC_{50} calculations are estimates, we should also obtain confidence intervals. The `drc` library provides a function called `ED` which will accomplish this. We can also ask what the survival was at 10% and 90% in addition to 50%, along with the confidence intervals for each.

At the R prompt type

```
ED(model.dose1,c(10,50,90), interval="delta")
```

And the output is shown below.

```
Estimated effective doses
(Delta method-based confidence interval(s))
     Estimate Std. Error   Lower   Upper
1:10  274.348     38.291 183.803  364.89
1:50  492.645     47.680 379.900  605.39
1:90  884.642    208.171 392.395 1376.89
```

Thus, the 95% confidence interval for the EC_{50} calculated from the four-parameter logistic curve was between the lower limit of 379.9 and upper limit of 605.39 mm NaCl.

EC_{50} from three-parameter model

Looking at the summary output from the four parameter logistic function, we see that the value for c was -0.085 and the p-value was 0.53, which suggests that the lower limit was not statistically different from zero. We would expect this given that the experiment had included a control of zero mm added salt. Thus, we can explore by how much the EC_{50} estimate changes when the additional parameter c is no longer estimated by calculating a **three parameter model** with `LL.3()` .

```
model.dose2 = drm(dose,fct=LL.3())
summary(model.dose2)
```

R output follows.

```
Model fitted: Log-logistic (ED50 as parameter) with lower limit at 0 (3 parms)
Parameter estimates:
               Estimate Std. Error   t-value p-value
b:(Intercept)   4.46194    0.76880   5.80378   4e-04
d:(Intercept)   1.00982    0.04866  20.75272   0e+00
e:(Intercept) 467.87842   25.24633  18.53253   0e+00
Residual standard error:
 0.08267671 (8 degrees of freedom)
```

The EC_{50} is the value of e: 467.88 mM NaCl.

How do the four- and three-parameter models compare? We can rephrase this as as statistical test of fit; which model fits the data better, a three-parameter or a four-parameter model?

At the R prompt type

```
anova(model.dose1, model.dose2)
```

The R output follows.

```
1st mode
fct:      LL.3()
2nd model
fct:      LL.4()
```

```
ANOVA table
          ModelDf      RSS Df F value p value
2nd model       8 0.054684
1st model       7 0.050012  1  0.6539  0.4453
```

Because the p-value is much greater than 5% we may conclude that the fit of the four-parameter model was not significantly better than the fit of the three-parameter model. Thus, based on our criteria we established in discussions of model fit in Chapters 16 – 18, we would conclude that the three-parameter model is the preferred model.

The plot below now includes the fit of the four-parameter model (red line) and the three-parameter model (green line) to the data (Fig. 20.10.4).

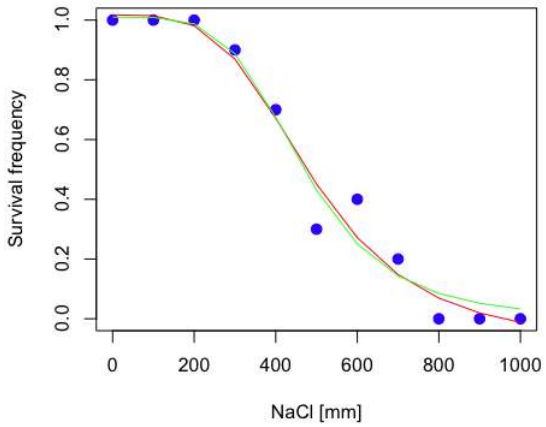


Figure 20.10.4: Four-parameter (red) and three-parameter (green) logistic models fitted to data.

The R command to make this addition to our active plot was

```
lines(dose,predict(model.dose2, data.frame(x=dose)),col="green")
```

We continue with our analysis of the three parameter model and produce the confidence intervals for the EC_{50} (modify the `ED()` statement above for `model.dose2` in place of `model.dose1`).

```
Estimated effective doses
(Delta method-based confidence interval(s))
     Estimate Std. Error   Lower  Upper
1:10  285.937     33.154 209.483 362.39
1:50  467.878     25.246 409.660 526.10
1:90  765.589     63.026 620.251 910.93
```

Thus, the 95% confidence interval for the EC_{50} calculated from the three-parameter logistic curve was between the lower limit of 409.7 and upper limit of 526.1 mm NaCl. The difference between upper and lower limits was 116.4 mm NaCl, a smaller difference than the interval calculated for the 95% confidence intervals from the four-parameter model (225.5 mm NaCl). This demonstrates the estimation trade-off: more parameters to estimate reduces the confidence in any one parameter estimate.

Additional notes of EC_{50} calculations

Care must be taken that the model fits the data well. What if we did not have observations throughout the range of the sigmoidal shape? We can explore this by taking a subset of the data.

```
dd = dose[1:6,]
```

Here, all values after dose 500 were dropped:

```
dd
  resp dose
1  1.0    0
2  1.0  100
3  1.0  200
4  0.9  300
5  0.7  400
6  0.3  500
```

and the plot does not show an obvious sigmoidal shape (Fig. 20.10.5).

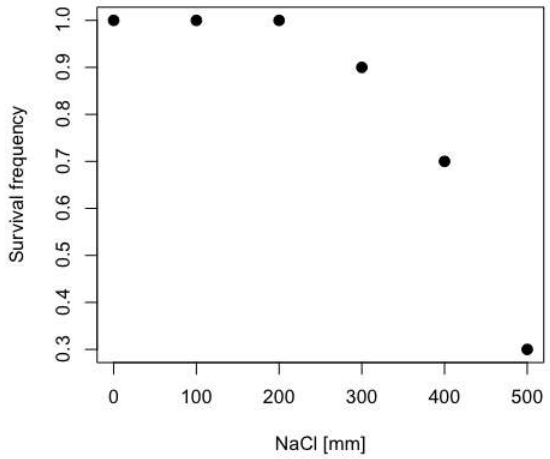


Figure 20.10.5: Plot of reduced data set.

We run the three-parameter model again, this time on the subset of the data.

```
model.dosedd = drm(dd,fct=LL.3())
summary(model.dosedd)
```

Output from the results are

```
Model fitted: Log-logistic (ED50 as parameter) with lower limit at 0 (3 parms)
Parameter estimates:
                Estimate Std. Error    t-value p-value
b:(Intercept)   6.989842   0.760112   9.195801  0.0027
d:(Intercept)   0.993391   0.014793  67.153883  0.0000
e:(Intercept) 446.882542   5.905728  75.669344  0.0000
Residual standard error:
 0.02574154 (3 degrees of freedom)
```

Conclusion? The estimate is different, but only just so, 447 vs. 468 mm NaCl.

Thus, within reason, the `drc` function performs well for the calculation of EC_{50}. Not all tools available to the student will do as well.

NLopt and nloptr

draft

Free open source library for nonlinear optimization.

Steven G. Johnson, The NLopt nonlinear-optimization package, http://github.com/stevengj/nlopt

https://cran.r-project.org/web/packages/nloptr/vignettes/nloptr.html

Alternatives to R

What about online tools? There are several online sites that will allow students to perform these kinds of calculations. Students familiar with MatLab know that it can be used to solve nonlinear equation problems. An open source alternative to MatLab is called GNU Octave, which can be installed on a personal computer or run online at http://octave-online.net. Students also may be aware of other sites, e.g., mycurvefit.com and IC50.tk.

Both of these free services performed well on the full dataset (results not shown), but fared poorly on the reduced subset: mycurvefit returned a value of 747.64 and IC50.tk returned an EC_{50} estimate of 103.6 (Michael Dohm, pers. obs.).

Simple inspection of the plotted values shows that these values are unreasonable.

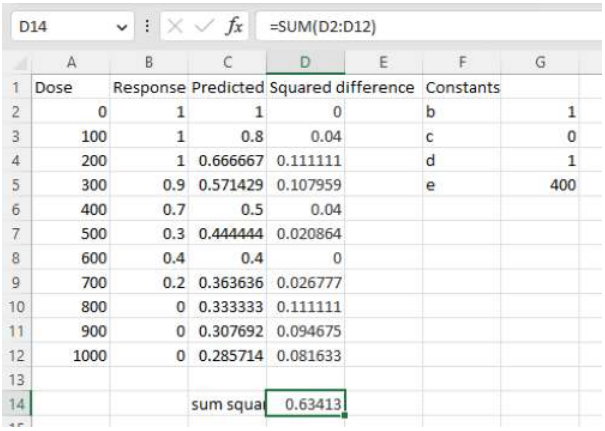
EC_{50} calculations with Microsoft Excel

Most versions of Microsoft Excel include an add-in called Solver, which will permit mathematical modeling. The add-in is not installed as part of the default installation, but can be installed via the Options tab in the File menu for a local installation or via Insert for the Microsoft Office online version of Excel (you need a free Microsoft account). The following instructions are for a local installation of Microsoft Office 365 and were modified from information provided by sharpstatistics.co.uk.

After opening Excel, set up worksheet as follows. Note that in order to use my formulas your spreadsheet values need to be set up exactly as I describe.

- Dose values in column A, beginning with row 2
- Response values in column B, beginning with row 2
- In column F type `b` in row 2, `c` in row 3, `d` in row 4, and `e` in row 5.
- In cells G2 – G5, enter the starting values. For this example, I used b = 1, c = 0, d = 1, e = 400.
  - For your own data you will have to explore use of different starting values.
- Enter headers in row 1.
  - Column A `Dose`
  - Column B `Response`
  - Column C `Predicted`
  - Column D `Squared difference`
  - Column F `Constants`
- In cell C14 enter `sum squares`
- In cell D14 enter `=sum(D2:D12)`

Here is an image of the worksheet (Fig. 20.10.6), with equations entered and values updated, but before running Solver.



	A	B	C	D	E	F	G
1	Dose	Response	Predicted	Squared difference		Constants	
2	0	1	1	0		b	1
3	100	1	0.8	0.04		c	0
4	200	1	0.666667	0.111111		d	1
5	300	0.9	0.571429	0.107959		e	400
6	400	0.7	0.5	0.04			
7	500	0.3	0.444444	0.020864			
8	600	0.4	0.4	0			
9	700	0.2	0.363636	0.026777			
10	800	0	0.333333	0.111111			
11	900	0	0.307692	0.094675			
12	1000	0	0.285714	0.081633			
13							
14			sum squa	0.63413			

Figure 20.10.6: Screenshot of Microsoft Excel worksheet containing our data set (col A & B), with formulas added and calculated. Starting values for constants in column G, rows 2 – 4.

Next, enter the functions.

- In cell C2 enter the four parameter logistic formula — type everything between the double quotes: " `=$G$3+(($G$4-$G$3)/(1+(A2/$G$5)^$G$2))` ". Next, copy or drag the formula to cell C12 to complete the predictions.
  - Note: for a three parameter model, replace the above formula with " `=(($G$4)/(1+(A2/$G$5)^$G$2))` ".
- In cell D2 type everything between the double quotes: " `=(B2-C2)^2` ".
- Next, copy or drag the formula to cell `D12` to complete the predictions.
- Now you are ready to run solver to estimate the values for b, c, d, and e.
  - Reminder: starting values must be in cells `G2:G5`
- Select cell `D14` and start solver by clicking on Data and looking to the right in the ribbon.

Solver is not normally installed as part of the default installation of Microsoft Excel 365. If the add-in has been installed you will see Solver in the Analyze box (Fig. 20.10.7).

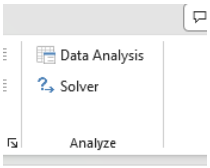


Figure 20.10.7: Screenshot of Microsoft Excel, Solver add-in available.

If Solver is not installed, go to **File** → **Options** → **Add-ins** (Fig. 20.10.8).

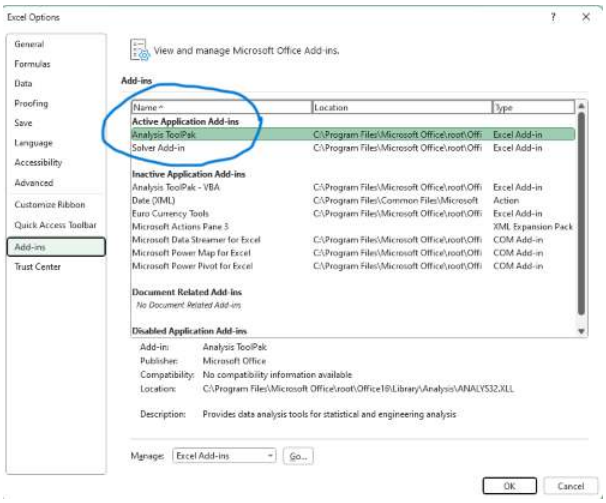


Figure 20.10.8: Screenshot of Microsoft Excel, Solver add-in available and ready for use.

Go to Microsoft support for assistance with add-ins.

With the spreadsheet completed and Solver available, click on Solver in the ribbon (Fig. 20.10.7) to begin. The screen shot from the first screen of Solver is shown below (Fig. 20.10.9).

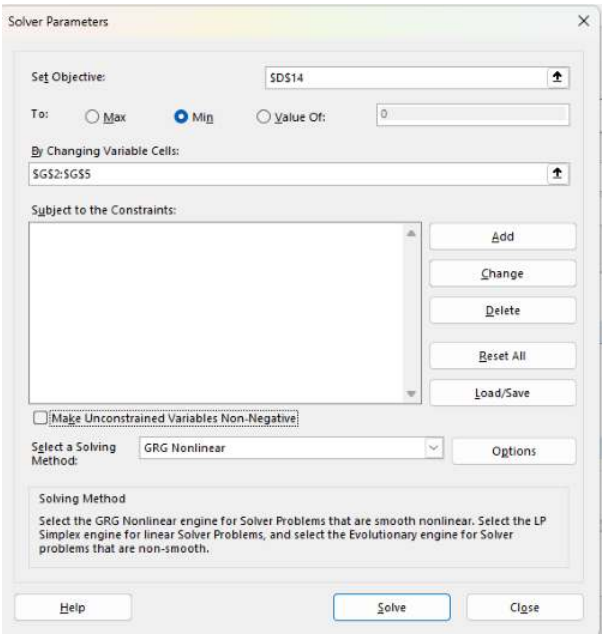


Figure 20.10.9: Screenshot of Microsoft Excel Solver menu.

Setup Solver

- **Set Objective**, enter the absolute cell reference to the sum squares value, `$D$14`
- Set **To:** Min.
- For **By Changing Variable Cells**, enter the range of cells for the four parameters, `$G$2:$G$5`
- Uncheck the box by "**Make Unconstrained Variables Non-Negative**."
  - Where constraints refers to any system of equalities or inequalities equations imposed on the algorithm.
- **Select a Solving method**, choose "GRG Nonlinear," the nonlinear programming solver option (not shown in Fig. 20.10.9 select by clicking on the down arrow).
- Click **Solve** button to proceed.

> **Note:**
>
> GRG Nonlinear is one of many optimization methods. In this case we calculate the minimum of the sum of squares — the difference between observed and predicted values from the logistic equation — given the range of observed values. GRG stands for Generalized Reduced Gradient and finds the local optima — in this case the minimum or valley — without any imposed constraints. See solver.com for additional discussion of this algorithm.

If all goes well, this next screen (Fig. 20.10.10) will appear, which shows the message "**Solver has converged to the current solution. All constraints are satisfied.**"

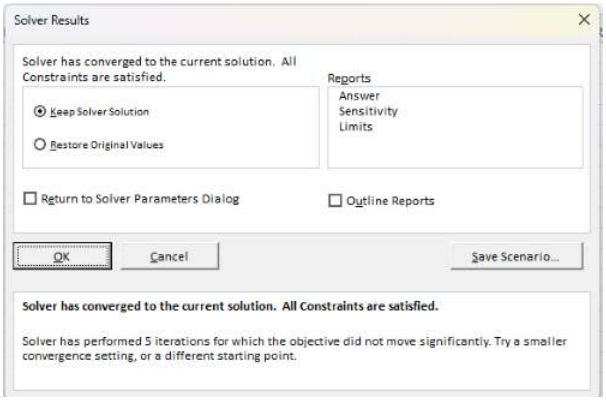


Figure 20.10.10: Screenshot showing solver has completed run.

Click OK and note that the values of the parameters have been updated (Table 20.10.1).

Table 20.10.1. Four-parameter logistic model, results from Solver.

Constant	Starting values	Values after solver
b	1	3.723008382
c	0	-0.088865487
d	1	1.017964505
e	400	493.9703594

Note the values obtained by Solver are virtually identical to the values obtained in the `drc` R package. The differences are probably because of the solver algorithm.

More interestingly, how did Solver do on the subset data set? Here are the results from a three-parameter logistic model (Table 20.10.2).

Table 20.10.2. Three parameter logistic model, results from Solver

Constant	Starting values	Full dataset, Solver results	Subset, Solver results
b	1	3.723008382	6.989855948
d	1	1.017964505	0.993391264
e	400	493.9703594	446.8819282

The results are again very close to results from the `drc` R package.

Thus, we would conclude that Solver and Microsoft Excel would be a reasonable choice for EC_{50} calculations, and much better than IC50.tk and mycurvefit.com. The advantage of R over Microsoft Excel is that the model building is more straightforward than the entering formulas in the cell reference format required by Excel.

Questions

[pending]

References and suggested readings

Beck B, Chen YF, Dere W, et al. Assay Operations for SAR Support. 2012 May 1 [Updated 2012 Oct 1]. In: Sittampalam GS, Coussens NP, Nelson H, et al., editors. *Assay Guidance Manual* [Internet]. Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004-. Available from: www.ncbi.nlm.nih.gov/books/NBK91994/

Di Veroli G. Y., Fornari C., Goldlust I., Mills G., Koh S. B., Bramhall J. L., Richards, F. M., Jodrell D. I. (2015) An automated fitting procedure and software for dose-response curves with multiphasic features. *Scientific Reports* 5: 14701.

Ritz, C., Baty, F., Streibig, J. C., Gerhard, D. (2015) Dose-Response Analysis Using R. *PLOS ONE*, 10(12), e0146021

Tjørve, K. M., & Tjørve, E. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. *PloS One*, 12(6), e0178691.

20.11: Plot a Newick tree

Introduction

The phrase “paradigm shift”, attributed to Kuhn (1962, see Wikipedia), may be well-worn and even abused today (Naughton 2012), but the shift in thinking from essential types and group thinking (essentialism) to viewing species as varying individuals in populations (populating thinking) revolutionized biology (O’Hara 1997, Sandvik 2008). **Tree thinking** is the manifestation of Charles Darwin’s “descent with modification” metaphor (Gregory 2008). Thus, every biology student should have ability to work with, and interpret, phylogenetic trees (tree thinking). The subject of creating and working with phylogenetic graphs is complicated with an extensive library. A good review is available from Holder and Lewis (2003) and readers should know Felsenstein’s book (2004).

Here, I include a modest, incomplete primer on working with trees in R.

- Loading the tree file
- Change tip names
- Write tip names to a text file
- Plot the tree as phylogram or cladogram
- Get node labels
- Re-root the tree
- Write a tree to a file

I assume that the student already has a set of species or other taxa; has gathered sequences (DNA or protein), aligned the sequences, and estimated a gene or phylogeny tree; and wishes to view and manipulate the tree in R. While these kinds of analyses can be done with R and R packages (see Task view: Phylogenetics), other software may be better choice for the student just beginning with phylogenetic tree building (see Unipro UGENE and MEGA, for examples). If the goal is just to view a tree file, or add annotations, then I recommend the iTOL tools.

Data formats

Phylogeny and gene trees are special cases of network graphs. Newick format (Wikipedia) is a common but limited representation of the tree which uses parentheses (groupings) and commas (branching). Other formats permit additional information; examples are Nexus file (Wikipedia) and the extension of Nexus to XML, NeXML (Wikipedia), and phyloXML (Wikipedia) formats. Our example uses Newick format.

Data set

I’ll use a “time tree” for an example. Tree from timetree.org, list of species (copy/paste list to a text file, load the text file Load list of of Species, then save the tree as a Newick file).

```
Alligator mississippiensis
Felis catus
Bos taurus
Gallus gallus
Pan troglodytes
Canis lupus
Homo sapiens
Anolis carolinensis
Macaca mulatta
Mus musculus
Didelphis virginiana
Sus scrofa
Oryctolagus cuniculus
Rattus norvegicus
```

R code

Requires the `ape` package. `Phylotools` and `Phytools` packages provide additional handy functions. References for these packages are listed at the end of this page.

```
require(ape)
require(phytools)
require(phylotools)
#If tree file, then
read.tree(file="tree14.nwk")
or
tree14 <- read.tree(file.choose())
#If no tree file saved, copy the Newick data use text="", replace example tree with y
tree14 <-read.tree(text="((Anolis_carolinensis:279.65697667,(Gallus_gallus:236.502662
(Didelphis_virginiana:158.59758758,
(((Felis_catus:54.32144118,Canis_lupus:54.32144118)'11':23.43351523,
(Bos_taurus:61.96598852,Sus_scrofa:61.96598852)'10':15.78896789)'19':18.70743276,
((Oryctolagus_cuniculus:82.14079889,
(Rattus_norvegicus:20.88741740,Mus_musculus:20.88741740)'9':61.25338149)'22':7.682388
(Macaca_mulatta:29.44154682,
(Pan_troglodytes:6.65090500,Homo_sapiens:6.65090500)'8':22.79064182)'6':60.38164060)'
#return information about the object
tree14
```

Output returned by R:

```
Phylogenetic tree with 14 tips and 13 internal nodes.
Tip labels:
Anolis_carolinensis, Gallus_gallus, Alligator_mississippiensis, Didelphis_virginiana,
Node labels:
, 13, 14, 27, 29, 19, ...
Rooted; includes branch lengths.
```

Change the tip names. Create a data frame with the tip labels and new tip names.

```
require(phylotools)
timeTreeTips <- tree14$tip.label
replaceTips <- c("Alligator", "Cat", "Chicken", "Chimpanzee", "Cow", "Dog", "Human",
"Lizard", "Macaque", "Mouse", "Opossum", "Pig", "Rabbit", "Rat")
myDat <- data.frame(timeTreeTips,replaceTips)
ntree14<- sub.taxa.label(tree14,myDat)
```

Collect and write the tip names to a text file

```
#Extract tips from newick file, write to text file
require(ape)
my.tips <- sort(tree14$tip.label)
#option 1
cat(my.tips,file="outfile.txt",sep="\n")
```

```
#option 2
my_conn = file("outfile.txt")
writeLines(my.tips,my_conn)
close(my_conn)
```

Next, make the plot.

```
plot(ntree14)
```

Result, a simple phylogram, i.e., a tree diagram with branching patterns and branch lengths proportional to amount of character change.

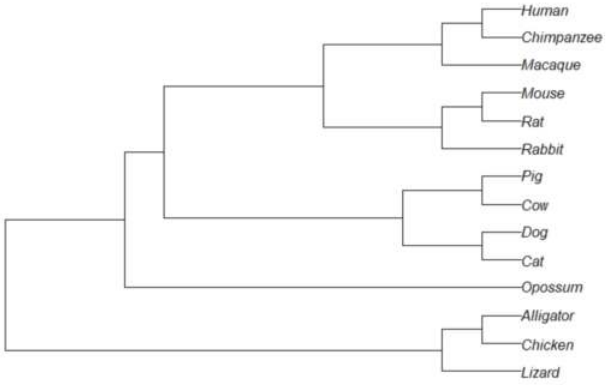


Figure 20.11.1: Phylogram plot of 14 taxa.

Or, change from default “phylogram” to “cladogram” view.

```
plot(tree14, type="cladogram")
```

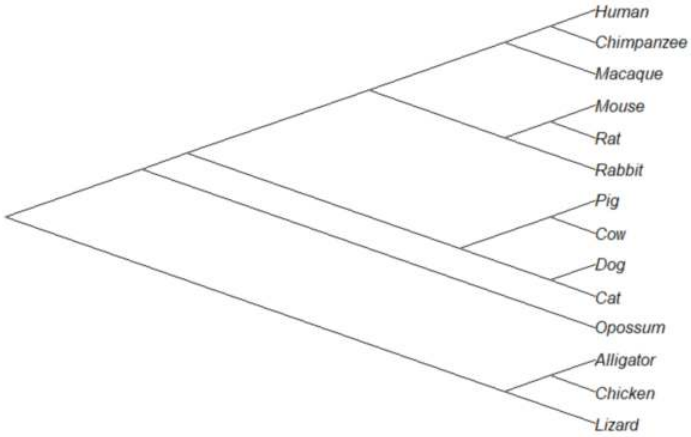


Figure 20.11.2: Cladogram view of the same 14 taxa.

Note that while the tree is rooted, it’s a midpoint rooting, the default setting in Newick files. For true root based on outgroup(s), identify the nodes, then select root.

Add node labels; `plot()` must be run first.

```
nodelabels()
```

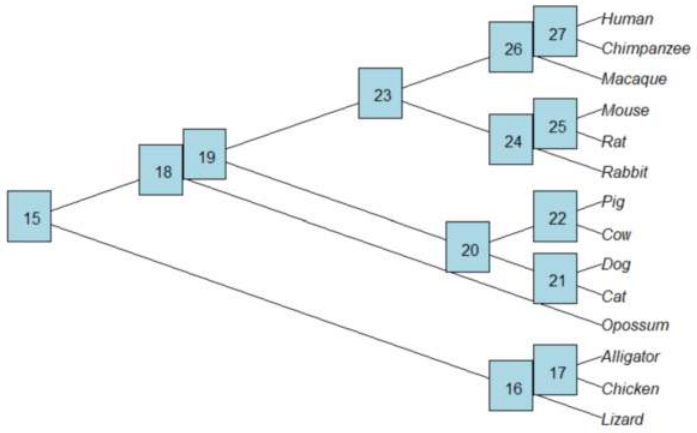


Figure 20.11.3: Plot of tree with labeled nodes.

The outgroup(s) were the reptiles (Alligator, Chicken, Lizard), so reroot at node 16.

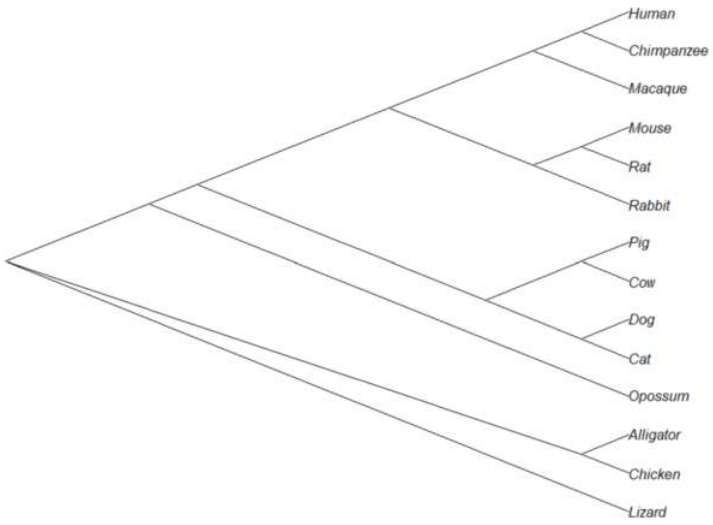


Figure 20.11.4: Re-rooted tree.

To write the tree to a file:

```
require(ape)
```

To export tree to Newick format

```
write.tree(tree14, file = "filename.nwk")
```

for Nexus format

```
write.nexus(tree14, file = "filename.nex")
```

Star phylogeny

Collapse the tree to a star phylogeny, an unlikely evolutionary model in which the species resulted from "… a single explosive adaptive radiation" (Felsenstein 1985). Star phylogeny is an extreme tree shape, or multifurcation (polytomy), where all tips derive from the same node (Colijn and Plazzotta 2018). This type of phylogeny can be viewed as a null model for inference (but see Bayesian "star phylogeny paradox," cf. Kolaczkowski and Thornton 2006).

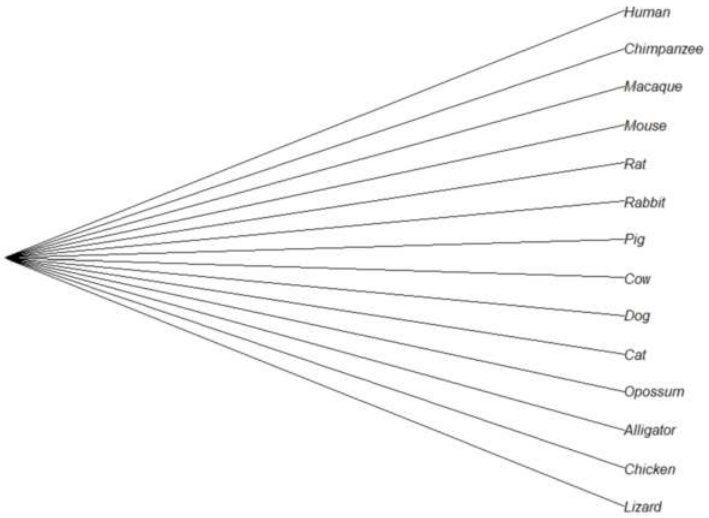


Figure 20.11.5: Star phylogeny.

Under a star phylogeny model, all taxa are assumed independent of each other, in contrast to the nested hierarchical model of evolution (e.g., Fig. 20.11.4), which shows a lack of independence among the taxa. More succinctly, comparisons fitted to uncorrected taxa may violate the assumption that errors are independent and identically distributed. Phylogenetically correct methods attempt to address the lack of independence among taxa for comparative analysis (Felsenstein 1985, Uyeda et al 2016). Biologists should know about Felsenstein's 1985 paper. Felsenstein's paper created a paradigm shift in how to analyze comparative datasets and has been cited more than ten thousand times (1 August 2023, Google Scholar). To put that number in context, the 1986 paper by Kary Mullis et al., which announced invention of PCR with thermally stable polymerase that has revolutionized molecular biology, has been cited 6721 times over that same period.

Additional packages of note

The R package `tanggle` works with the package `ggtree` and advantage of the ggplot2 environment. Contains many functions to work with phylogeny graphs including re-rooting and swapping nodes. The package is available from Bioconductor,

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("tanggle")
```

`ggtree` is also a Bioconductor package, not available at CRAN.

Online viewers

Many browser-based tree viewers are available online, including icytree.org and iTOL tools. Additional tree viewers listed at Wikipedia.

References and suggested readings

Colijn C, Plazzotta G. (2018). A Metric on Phylogenetic Tree Shapes. *Syst Biol*. 67(1):113-126. doi: 10.1093/sysbio/syx046. PMID: 28472435; PMCID: PMC5790134.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist* 125(1):1-15.

Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland, MA: Sinauer associates.

Gregory, T. R. (2008). Understanding evolutionary trees. *Evolution: Education and Outreach*, *1*(2), 121-137.

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews genetics*, *4*(4), 275-284.

Kolaczkowski, B., & Thornton, J. W. (2006). Is There a Star Tree Paradox? *Molecular Biology and Evolution*, *23*(10), 1819–1823. https://doi.org/10.1093/molbev/msl059

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press: Chicago.

Mullis, K., Faloona, F., Scharf, S., Saiki, R. K., Horn, G. T., & Erlich, H. (1986, January). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 51, pp. 263-273). Cold Spring Harbor Laboratory Press.

Naughton, J. (2012, August 18). *Thomas Kuhn: The man who changed the way the world looked at science*. The Guardian. https://www.theguardian.com/science/...ic-revolutions

O'Hara, R. J. (1997). Population thinking and tree thinking in systematics. *Zoologica scripta*, *26*(4), 323-329.

Paradis, E. (2012) *Analysis of Phylogenetics and Evolution with R* (Second Edition). New York: Springer.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528.

Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3, 217-223.

Sandvik, H. (2008). Tree thinking cannot taken for granted: challenges for teaching phylogenetics. *Theory in Biosciences*, *127*(1), 45-51.

Uyeda, J. C., Zenil-Ferguson, R., & Pennell, M. W. (2018). Rethinking phylogenetic comparative methods. *Systematic Biology*, *67*(6), 1091-1109.

Zhang, J., Pei, N., & Mi, X. (2012). phylotools: Phylogenetic tools for Eco-phylogenetics. R package version 0.1, 2.

20.12: Phylogenetically independent contrasts

Introduction

Assumption of independence among the subjects in a study is a key assumption. Comparisons among species are a common experimental approach in evolutionary biology. Typical statistical approaches include use of ANOVA or linear regression approaches. A basic assumption of ANOVA is that sampling units are independent (13.1 – ANOVA assumptions). Prior to the 1980s, it was rarely appreciated in comparative analysis that species are not independent sample units (Harvey and Pagel 1991); evolution produced nested hierarchical relationships among the species. We recognize this with phylogenies (Felsenstein 1985, Harvey and Pagel 1991, Martins 1996, Garland et al 2005). Mice and rats share a more recent common ancestor, and cattle and pigs share a more recent common ancestor, than do mice and cattle, for example. Felsenstein (1985, 1988) is largely credited for making the argument that Type I error is likely if phylogeny is ignored, and, importantly, provided an algorithm: **Phylogenetic Independent Contrasts, PIC**, which provided a simple way to correct for phylogenetic nonindependence. Felsentein's landmark 1985 paper has been cited more than ten thousand times (Feb 2024). However, like most innovations, PIC should not be blindly applied in all comparative analysis (e.g., unreplicated evolutionary events, Uyeda et al 2018).

Logic of PIC

Treating comparative data, e.g., species, as a collection of independent samples implies that the evolutionary history was a spontaneous burst, or star-like phylogeny.

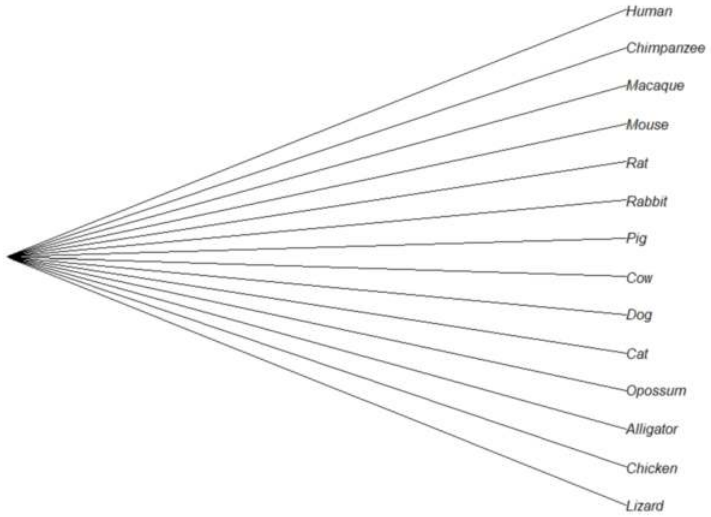


Figure 20.12.1: Star phylogeny (same image shown in Figure 20.11.5.

But what nature provides is nonindependence (Fig. 20.12.2 for more about star phylogeny in PIC see discussion in Garland et al 2005), which should be accounted for during statistical analysis.

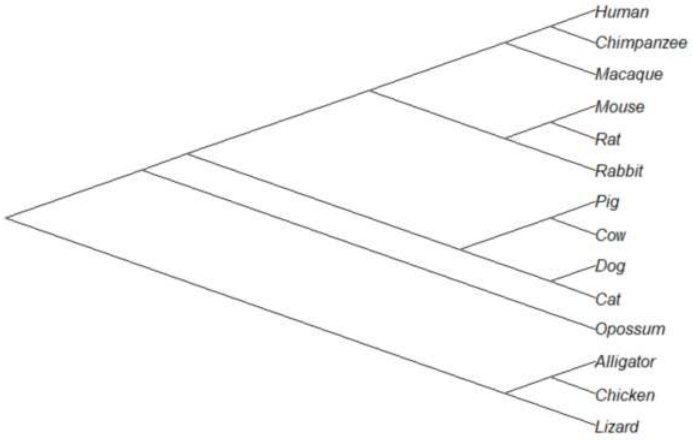


Figure 20.12.2: A cladogram for same species, showing the hierarchical, nested relationships among taxa, what nature actually provides (same image shown Figure 20.11.2.

R package, `phytools` , `ape`

Lots of good references on this important subject. For now, see

Chapter 4.2, Estimating rates using independent contrasts, by Dr Luke Harmon

and a tutorial from same author, available at

https://lukejharmon.github.io/ilhabela/instruction/2015/07/02/phylogenetic-independent-contrasts/

Questions

[pending]

References and suggested readings

Felsenstein, J (1985) Phylogenies and the comparative method. *American Naturalist* 125(1):1-15.

Felsenstein, J. (1988) Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19, 445–471.

Garland Jr, T., Bennett, A. F., & Rezende, E. L. (2005). Phylogenetic approaches in comparative physiology. *Journal of experimental Biology*, *208*(16), 3015-3035.

Harvey, P.H. and Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, Oxford Series in Ecology and Evolution.

Martins, E.P. (1996) *Phylogenies and the Comparative Method in Animal Behavior*. Oxford University Press.

Paradis, E. (2012) *Analysis of Phylogenetics and Evolution with R* (Second Edition). New York: Springer.

Paradis, E. and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528.

Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3, 217-223.

Uyeda, J. C., Zenil-Ferguson, R., & Pennell, M. W. (2018). Rethinking phylogenetic comparative methods. *Systematic Biology*, *67*(6), 1091-1109.

Zhang, J., Pei, N., & Mi, X. (2012). phylotools: Phylogenetic tools for Eco-phylogenetics. R package version 0.1, 2.

20.13: How to get the distances from a distance tree

Introduction

Extract the **patristic distance**, the sum of the branch lengths that link two nodes in a tree, for each pair of species.

This **distance** — see our Chapter 16.6 – Similarity and Distance— is the proportion (p) of amino acid (or nucleotide for DNA or RNA) sites at which the two sequences to be compared are different. It is obtained by dividing the number of amino acid differences by the total number of sites compared. It does not make any correction for multiple substitutions at the same site or differences in evolutionary rates among sites. On a **gene tree** (Fig. 20.13.1), distances are the lengths of the branches connecting the taxa. We want to know, how different are two species for the given protein? That's the distance between them in proportion of amino acid sites that are different by total number compared.

Example

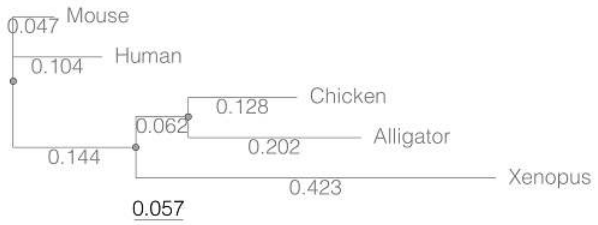


Figure 20.13.1: A gene tree of the product (protein HBA1) with five species.

Here's the Newick format for the tree (HBA1.nwk)

```
(Mouse:0.0474516,Human:0.104063,((Chicken:0.127652,Alligator:0.202421):0.0616593,Xeno
```

R code to extract distances and output sorted, **pairwise comparisons** to a text file:

```
library(ape)

# Create a function
getDis <- function(tree, tips) {
     myTree <- cophenetic(tree)
     myTree <- myTree[,tips]
     xy <- t(combn(colnames(myTree), 2))
     xy <- xy[order(xy[,1], xy[,2]),]
     myOut <- data.frame(xy, myTree[xy])
     colnames(myOut) <- c("Spp1", "Spp2", "Distance")
   return(myOut)
}

# Read a tree file, Newick format
tree5 <-read.tree(text="(Mouse:0.0474516,Human:0.104063,((Chicken:0.127652,Alligator:

# get taxa names from the tree file
all.tips <- tree5$tip.label; all.tips

# Run the function
myDis <- getDis(tree5, all.tips)
```

```
# Check the output
head(myDis)

# Create the results file
write.csv(myDis, file = "my_out.txt")
```

Example output from `head(myDist)`

```
       Spp1    Spp2  Distance
1 Alligator Xenopus 0.6868813
2 Chicken Alligator 0.3300730
3 Chicken   Xenopus 0.6121123
4 Human   Alligator 0.5120823
5 Human     Chicken 0.4373133
6 Human     Xenopus 0.6708030
```

The function sorts first by Spp1, then by Spp2.

Molecular clock plot

Collect divergence times from timetree.org

Spp1	Spp2	Time (median MYA)
Alligator	Xenopus	352
Chicken	Alligator	245
Chicken	Xenopus	352
Human	Alligator	319
Human	Chicken	319
Human	Xenopus	352

A scatterplot of distance HBA protein sequence by $\log_{10}$-transformed millions of years ago divergence time is shown in Figure 20.13.2 Note that, although tempting, calculating the slope from a linear regression to estimate the rate of evolution would not be appropriate without accounting for the lack of independence of the data (see Phylogenetically independent contrasts). Better methods exist, including calculating rate of change after fitting a model that assumes a strict clock vs relaxed clock.

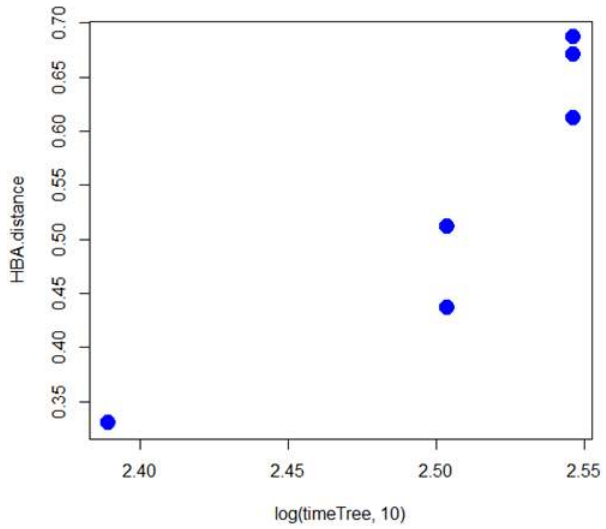


Figure 20.13.2: Scatterplot of HBA distance by $\log_{10}$(MYA) divergence time

Questions

[pending]

Suggested readings

Bevan, R. B., Lang, B. F., & Bryant, D. (2005). Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Systematic biology*, *54*(6), 900-915.

20.14: Binary classification

Future home for

Prediction

Linear discriminant analysis

Machine learning

Supervised learning

Training data

References

Li, J. J., & Tong, X. (2020). Statistical hypothesis testing versus machine learning binary classification: Distinctions and guidelines. *Patterns*, 1(7).

CHAPTER OVERVIEW

Appendix

A.1: Distribution tables

Tables of common probability distributions

The appendix provides critical values and probabilities for a few of the most common probability distributions. The tables were generated by appropriate functions in R. Code is provides with each statistical table

Table of Z of standard normal probabilities

Table of Chi-square critical values

Table of Critical values of Student's t-distribution

Table of Critical values of F-distribution

Interpolating p-values

We have a calculated test statistic of 3.333 from a chi-square test; how likely is it that our test statistic value of 3.333 and the null hypothesis are true? (Remember, "true" in this case is a shorthand for our data was sampled from, for example, a population in which the Hardy-Weinberg expectations hold). When I check the table of critical values of the chi-square test for the "exact" p-value, I find that our test statistic value falls between a p-value of 0.10 and 0.05 (represented in the table below). How can I find our exact p-value, u (unknown)?

statistic	**p-value**
3.841	0.05
3.333	u
2.706	0.10

Short answer, use R. In the case of interpolating to find u. If we assume the change in probability between 2.706 and 3.841 for the chi-square distribution is linear (it's not, but it's close), then we can do so simple interpolation.

We set up what we know on the right hand side, equal to what we don't know on the left hand side of the equation:

$$\frac{u-0.10}{0.05-0.10}=\frac{3.333-2.706}{3.841-2.706}$$

and solve for u. Then, u is equal to 0.0724.

R function `pchisq()` gives a value of P = 0.0679. Our interpolated value is close, but not the same. Of course, you should go with the result from R; we mention how to get the approximate p-value by interpolation for completeness, and, in some rare instances, you might need to make the calculation.

A.2: Table of Z of standard normal probabilities

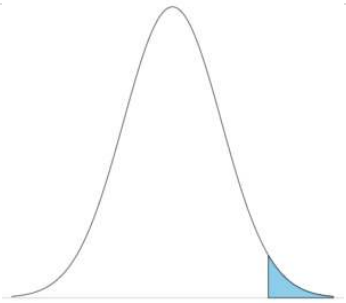


Figure $A.2.1$.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.468119	0.464144
0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.428576	0.424655
0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.389739	0.385908
0.3	0.382089	0.378281	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.351973	0.348268
0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.315614	0.312067
0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.280957	0.277595
0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.248252	0.245097
0.7	0.241964	0.238852	0.235763	0.232695	0.229650	0.226627	0.223627	0.220650	0.217695	0.214764
0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.189430	0.186733
0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.163543	0.161087
1.0	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.140071	0.137857
1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.119000	0.117023
1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.100273	0.098525
1.3	0.096800	0.095098	0.093418	0.091759	0.090123	0.088508	0.086915	0.085343	0.083793	0.082264
1.4	0.080757	0.079270	0.077804	0.076359	0.074934	0.073529	0.072145	0.070781	0.069437	0.068112
1.5	0.066807	0.065522	0.064255	0.063008	0.061780	0.060571	0.059380	0.058208	0.057053	0.055917
1.6	0.054799	0.053699	0.052616	0.051551	0.050503	0.049471	0.048457	0.047460	0.046479	0.045514
1.7	0.044565	0.043633	0.042716	0.041815	0.040930	0.040059	0.039204	0.038364	0.037538	0.036727
1.8	0.035930	0.035148	0.034380	0.033625	0.032884	0.032157	0.031443	0.030742	0.030054	0.029379
1.9	0.028717	0.028067	0.027429	0.026803	0.026190	0.025588	0.024998	0.024419	0.023852	0.023295
2.0	0.022750	0.022216	0.021692	0.021178	0.020675	0.020182	0.019699	0.019226	0.018763	0.018309
2.1	0.017864	0.017429	0.017003	0.016586	0.016177	0.015778	0.015386	0.015003	0.014629	0.014262
2.2	0.013903	0.013553	0.013209	0.012874	0.012545	0.012224	0.011911	0.011604	0.011304	0.011011
2.3	0.010724	0.010444	0.010170	0.009903	0.009642	0.009387	0.009137	0.008894	0.008656	0.008424
2.4	0.008198	0.007976	0.007760	0.007549	0.007344	0.007143	0.006947	0.006756	0.006569	0.006387
2.5	0.006210	0.006037	0.005868	0.005703	0.005543	0.005386	0.005234	0.005085	0.004940	0.004799
2.6	0.004661	0.004527	0.004396	0.004269	0.004145	0.004025	0.003907	0.003793	0.003681	0.003573
2.7	0.003467	0.003364	0.003264	0.003167	0.003072	0.002980	0.002890	0.002803	0.002718	0.002635

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.8	0.002555	0.002477	0.002401	0.002327	0.002256	0.002186	0.002118	0.002052	0.001988	0.001926
2.9	0.001866	0.001807	0.001750	0.001695	0.001641	0.001589	0.001538	0.001489	0.001441	0.001395
3.0	0.001350	0.001306	0.001264	0.001223	0.001183	0.001144	0.001107	0.001070	0.001035	0.001001
3.1	0.000968	0.000935	0.000904	0.000874	0.000845	0.000816	0.000789	0.000762	0.000736	0.000711
3.2	0.000687	0.000664	0.000641	0.000619	0.000598	0.000577	0.000557	0.000538	0.000519	0.000501
3.3	0.000483	0.000466	0.000450	0.000434	0.000419	0.000404	0.000390	0.000376	0.000362	0.000349
3.4	0.000337	0.000325	0.000313	0.000302	0.000291	0.000280	0.000270	0.000260	0.000251	0.000242
3.5	0.000233	0.000224	0.000216	0.000208	0.000200	0.000193	0.000185	0.000178	0.000172	0.000165
3.6	0.000159	0.000153	0.000147	0.000142	0.000136	0.000131	0.000126	0.000121	0.000117	0.000112
3.7	0.000108	0.000104	0.000100	0.000096	0.000092	0.000088	0.000085	0.000082	0.000078	0.000075
3.8	0.000072	0.000069	0.000067	0.000064	0.000062	0.000059	0.000057	0.000054	0.000052	0.000050
3.9	0.000048	0.000046	0.000044	0.000042	0.000041	0.000039	0.000037	0.000036	0.000034	0.000033
4.0	0.000032	0.000030	0.000029	0.000028	0.000027	0.000026	0.000025	0.000024	0.000023	0.000022

where standard refers to mean $\mu = 0$ and standard deviation $\sigma = 1$

$$Z = \frac{X_i - \mu}{\sigma}$$

R command

```
Z <- seq(-4,4, by=.1)
pnorm(c(Z), mean=0, sd=1, lower.tail=FALSE)
```

A.3: Table of Chi-square critical values

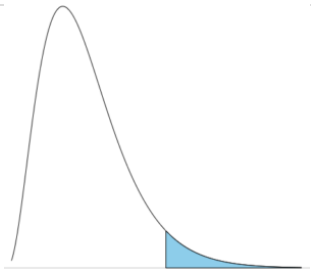


Figure $A.3.1$.

$\alpha(1)$	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
DF/1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.108	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.730
4	5.385	7.779	9.488	11.143	13.277	14.860	16.424	18.467	19.997
5	6.626	9.236	11.070	12.833	15.086	16.750	18.386	20.515	22.105
6	7.841	10.645	12.592	14.449	16.812	18.548	20.249	22.458	24.103
7	9.037	12.017	14.067	16.013	18.475	20.278	22.040	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.868
9	11.389	14.684	16.919	19.023	21.666	23.589	25.462	27.877	29.666
10	12.549	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.420
11	13.701	17.275	19.675	21.920	24.725	26.757	28.729	31.264	33.137
12	14.845	18.549	21.026	23.337	26.217	28.300	30.318	32.909	34.821
13	15.984	19.812	22.362	24.736	27.688	29.819	31.883	34.528	36.478
14	17.117	21.064	23.685	26.119	29.141	31.319	33.426	36.123	38.109
15	18.245	22.307	24.996	27.488	30.578	32.801	34.950	37.697	39.719
16	19.369	23.542	26.296	28.845	32.000	34.267	36.456	39.252	41.308
17	20.489	24.769	27.587	30.191	33.409	35.718	37.946	40.790	42.879
18	21.605	25.989	28.869	31.526	34.805	37.156	39.422	42.312	44.434
19	22.718	27.204	30.144	32.852	36.191	38.582	40.885	43.820	45.973
20	23.828	28.412	31.410	34.170	37.566	39.997	42.336	45.315	47.498
21	24.935	29.615	32.671	35.479	38.932	41.401	43.775	46.797	49.011
22	26.039	30.813	33.924	36.781	40.289	42.796	45.204	48.268	50.511
23	27.141	32.007	35.172	38.076	41.638	44.181	46.623	49.728	52.000
24	28.241	33.196	36.415	39.364	42.980	45.559	48.034	51.179	53.479
25	29.339	34.382	37.652	40.646	44.314	46.928	49.435	52.620	54.947
26	30.435	35.563	38.885	41.923	45.642	48.290	50.829	54.052	56.407
27	31.528	36.741	40.113	43.195	46.963	49.645	52.215	55.476	57.858
28	32.620	37.916	41.337	44.461	48.278	50.993	53.594	56.892	59.300
29	33.711	39.087	42.557	45.722	49.588	52.336	54.967	58.301	60.735
30	34.800	40.256	43.773	46.979	50.892	53.672	56.332	59.703	62.162
35	40.223	46.059	49.802	53.203	57.342	60.275	63.076	66.619	69.199
40	45.616	51.805	55.758	59.342	63.691	66.766	69.699	73.402	76.095

$\alpha(1)$	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
50	56.334	63.167	67.505	71.420	76.154	79.490	82.664	86.661	89.561

R command

```
qchisq(c(alpha), df=df, lower.tail=FALSE)
```

where `alpha` is one-tailed probability, `df` is number of degrees of freedom, and `lower.tail=FALSE` means each cell is to be read as equal to or greater than the critical value.

A.4: Table of critical values of Student's t-distribution

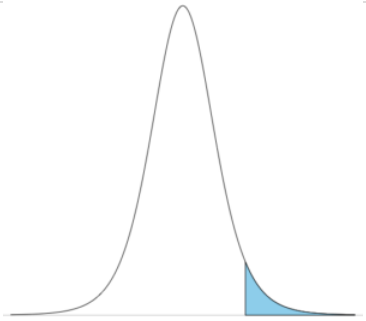


Figure $A.4.1$.

α (1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2)	0.50	0.2	0.10	0.05	0.02	0.01	0.005	0.002	0.001
DF/1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959

α (1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2)	0.50	0.2	0.10	0.05	0.02	0.01	0.005	0.002	0.001
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015

α (1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2)	0.50	0.2	0.10	0.05	0.02	0.01	0.005	0.002	0.001
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707

α (1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2)	0.50	0.2	0.10	0.05	0.02	0.01	0.005	0.002	0.001
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

Note: Here's the table at Wikipedia (Links to an external site.).

R command used to generate this table:

```
qt(c(alpha), df=df, lower.tail=FALSE)
```

where `alpha` is one-tailed probability, `df` is number of degrees of freedom, and `lower.tail=FALSE` means each cell is to be read as equal to or greater than the critical value.

A.5: Table of critical values of F-distribution

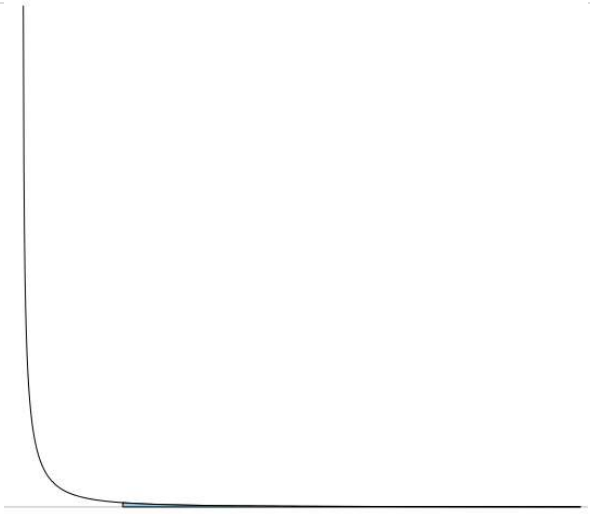


Figure $A.5.1.$

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001
$\alpha(2)$	0.5	0.200	0.1	0.05	0.02	0.01	0.005	0.002
Df=1,Df=1	5.828427	39.863	161.447639	647.789011	4052.180695	16210.72272	64844.89087	405284.0679
2	2.571	8.526	18.512821	38.506329	98.502513	198.501253	398.50063	998.50025
3	2.024	5.538	10.127964	17.443443	34.116222	55.551957	89.58433	167.02922
4	1.807	4.545	7.708647	12.217863	21.19769	31.332772	45.67398	74.13729
5	1.692	4.060	6.607891	10.006982	16.258177	22.784781	31.40667	47.18078
6	1.621	3.776	5.987378	8.813101	13.745023	18.634996	24.80731	35.50749
7	1.573	3.589	5.591448	8.072669	12.246383	16.235558	21.1107	29.24519
8	1.538	3.458	5.317655	7.570882	11.258624	14.688199	18.77965	25.41476
9	1.512	3.360	5.117355	7.209283	10.561431	13.613609	17.18757	22.85713
10	1.491	3.285	4.964603	6.936728	10.044289	12.82647	16.03626	21.0396
11	1.475	3.225	4.844336	6.72413	9.646034	12.226311	15.16738	19.68679
12	1.461	3.177	4.747225	6.553769	9.330212	11.75423	14.48958	18.64332
13	1.450	3.136	4.667193	6.414254	9.073806	11.37354	13.94676	17.81542
14	1.440	3.102213	4.60011	6.297939	8.861593	11.060253	13.50264	17.14336
15	1.432	3.073186	4.543077	6.199501	8.683117	10.798049	13.13278	16.58742
20	1.404	2.974653	4.351244	5.871494	8.095958	9.943935	11.94005	14.81878
30	1.376	2.880695	4.170877	5.567535	7.562476	9.179677	10.8893	13.29301

Numerator Df = 2

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001
$\alpha(2)$	0.5	0.200	0.1	0.05	0.02	0.01	0.005	0.002
Df=2,Df=1	7.500	49.500	199.500	799.500	4999.500	19999.500	79999.500	499999.500
2	3.000	9.000	19.000	39.000	99.000	199.000	399.000	999.000
3	2.280	5.462	9.552	16.044	30.817	49.799	79.933	148.500

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001
$\alpha(2)$	0.5	0.200	0.1	0.05	0.02	0.01	0.005	0.002
4	2.000	4.325	6.944	10.649	18.000	26.284	38.000	61.246
5	1.853	3.780	5.786	8.434	13.274	18.314	24.964	37.122
6	1.762	3.463	5.143	7.260	10.925	14.544	19.104	27.000
7	1.701	3.257	4.737	6.542	9.547	12.404	15.887	21.689
8	1.657	3.113	4.459	6.059	8.649	11.042	13.889	18.494
9	1.624	3.006	4.256	5.715	8.022	10.107	12.539	16.387
10	1.598	2.924	4.103	5.456	7.559	9.427	11.572	14.905
11	1.577	2.860	3.982	5.256	7.206	8.912	10.848	13.812
12	1.560	2.807	3.885	5.096	6.927	8.510	10.287	12.974
13	1.545	2.763	3.806	4.965	6.701	8.186	9.839	12.313
14	1.533	2.726	3.739	4.857	6.515	7.922	9.475	11.779
15	1.523	2.695	3.682	4.765	6.359	7.701	9.173	11.339
20	1.487	2.589	3.493	4.461	5.849	6.986	8.206	9.953
30	1.452	2.489	3.316	4.182	5.390	6.355	7.365	8.773

Numerator Df = 3

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001
$\alpha(2)$	0.5	0.200	0.1	0.05	0.02	0.01	0.005	0.002
Df=3,Df=1	8.200	53.593	215.707	864.163	5403.352	21614.741	86460.299	540379.200
2	3.153	9.162	19.164	39.165	99.166	199.166	399.167	999.167
3	2.356	5.391	9.277	15.439	29.457	47.467	76.056	141.109
4	2.047	4.191	6.591	9.979	16.694	24.259	34.956	56.177
5	1.884	3.619	5.409	7.764	12.060	16.530	22.426	33.202
6	1.784	3.289	4.757	6.599	9.780	12.917	16.867	23.703
7	1.717	3.074	4.347	5.890	8.451	10.882	13.843	18.772
8	1.668	2.924	4.066	5.416	7.591	9.596	11.979	15.829
9	1.632	2.813	3.863	5.078	6.992	8.717	10.726	13.902
10	1.603	2.728	3.708	4.826	6.552	8.081	9.833	12.553
11	1.580	2.660	3.587	4.630	6.217	7.600	9.167	11.561
12	1.561	2.606	3.490	4.474	5.953	7.226	8.652	10.804
13	1.545	2.560	3.411	4.347	5.739	6.926	8.242	10.209
14	1.532	2.522	3.344	4.242	5.564	6.680	7.910	9.729
15	1.520	2.490	3.287	4.153	5.417	6.476	7.634	9.335
20	1.481	2.380	3.098	3.859	4.938	5.818	6.757	8.098
30	1.443	2.276	2.922	3.589	4.510	5.239	5.999	7.054

Numerator Df = 4

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001
Df=4,Df=1	8.581	55.833	224.583	899.583	5624.583	22499.583	89999.583	562499.600
$\alpha(2)$	0.5	0.200	0.1	0.05	0.02	0.01	0.005	0.002
2	3.282	9.243	19.247	39.248	99.249	199.250	399.250	999.250
3	2.390	5.343	9.117	15.101	28.710	46.195	73.948	137.100
4	2.064	4.107	6.388	9.605	15.977	23.155	33.303	53.436
5	1.893	3.520	5.192	7.388	11.392	15.556	21.048	31.085
6	1.787	3.181	4.534	6.227	9.148	12.028	15.652	21.924
7	1.716	2.961	4.120	5.523	7.847	10.050	12.733	17.198
8	1.664	2.806	3.838	5.053	7.006	8.805	10.941	14.392
9	1.625	2.693	3.633	4.718	6.422	7.956	9.741	12.560
10	1.595	2.605	3.478	4.468	5.994	7.343	8.888	11.283
11	1.570	2.536	3.357	4.275	5.668	6.881	8.252	10.346
12	1.550	2.480	3.259	4.121	5.412	6.521	7.762	9.633
13	1.534	2.434	3.179	3.996	5.205	6.233	7.373	9.073
14	1.519	2.395	3.112	3.892	5.035	5.998	7.057	8.622
15	1.507	2.361	3.056	3.804	4.893	5.803	6.796	8.253
20	1.465	2.249	2.866	3.515	4.431	5.174	5.967	7.096
30	1.424	2.142	2.690	3.250	4.018	4.623	5.253	6.125

Numerator Df = 5

$\alpha(1)$	0.25	0.100	0.05	0.025	0.01	0.005	0.0025	0.001
$\alpha(2)$	0.5	0.200	0.1	0.05	0.02	0.01	0.005	0.002
Df=5,Df=1	8.820	57.240	230.162	921.848	5763.650	23055.798	92224.393	576404.600
2	3.280	9.293	19.296	39.298	99.299	199.300	399.300	999.300
3	2.409	5.309	9.013	14.885	28.237	45.392	72.621	134.580
4	2.072	4.051	6.256	9.364	15.522	22.456	32.261	51.712
5	1.895	3.453	5.050	7.146	10.967	14.940	20.178	29.752
6	1.785	3.108	4.387	5.988	8.746	11.464	14.884	20.803
7	1.711	2.883	3.972	5.285	7.460	9.522	12.031	16.206
8	1.658	2.726	3.687	4.817	6.632	8.302	10.283	13.485
9	1.617	2.611	3.482	4.484	6.057	7.471	9.116	11.714
10	1.585	2.522	3.326	4.236	5.636	6.872	8.288	10.481
11	1.560	2.451	3.204	4.044	5.316	6.422	7.671	9.578
12	1.539	2.394	3.106	3.891	5.064	6.071	7.196	8.892
13	1.521	2.347	3.025	3.767	4.862	5.791	6.820	8.354
14	1.507	2.307	2.958	3.663	4.695	5.562	6.515	7.922
15	1.494	2.273	2.901	3.576	4.556	5.372	6.263	7.567
20	1.450	2.158	2.711	3.289	4.103	4.762	5.463	6.461
30	1.407	2.049	2.534	3.026	3.699	4.228	4.776	5.534

R code used was

```
qf(c(alpha), df1=dfn, df2=dfd, lower.tail=FALSE)
```

Where `dfn` refers to numerator degrees of freedom and `dfd` refers to denominator degrees of freedom.

A.6: Install R

This page presents a detailed guide about how to install R onto your computer (LDE). Additional install R help was provided in Chapter 1.1 – A quick look at R and R Commander. Instructions for RStudio are also provided (optional for BI311 students). A guide to install R Commander is provided in Install R Commander. Instructions for how to run R via a "cloud computing" (serverless) option are also provided, Use R in the Cloud. For help upgrading installed packages after upgrading new R version, see R packages.

Note:

Installation guides quickly become outdated. This page was created first in September 2019 and last updated 25 January 2024 and describes working installation protocols at that time. As of October 2023, R-4.3.2 was current version. Instructions for Win10 and Win11 are the same. Instructions for Intel-based macOS are the same; with Apple's switch to ARM64 (M1, M2), changes have been made. Going forward, the instructions on this page, but not my videos — version numbers need to be updated in the videos, are likely to be the same for new R versions. Per usual caveat that my advice is offered for instructional purposes and in no way implies warranty against damage or guarantee of success.

Run R on your computer (i.e., local development environment or LDE)

1. **Windows PCs**, download the base application from https://cran.r-project.org, select Download R for Windows, and install the R software as you would any other software. All of you are likely to have the 64-bit version of Windows 11, so install the 64-bit version of R. Follow the instructions as they are presented. Screenshots of the install process are available at the end of this page (click here or scroll down to **Win11 setup, Screenshots**).
    - Current versions of Microsoft Windows come in several flavors, the simplest distinction is between home and pro. R runs perfectly well on both.
        - Windows 10 is reaching end of life cycle.
        - Some inexpensive Microsoft Windows PCs are built on ARM64, not Intel or AMD64 CPU. Thus, installing R and or RStudio may prove problematic.
    - You should install R with Administrator privileges. Highlight the install file, right-click the file, and select "Run as administrator" from the popup menu.
    - When you first try to run R you may get a popup screen "Windows protected your PC," locate and click on the "More info" link and select "Run anyway."
        - This in no way will harm your computer — provided you have downloaded from official sites. R is a verified program. Microsoft has taken an aggressive line on developers and favors apps that are part of their app store.
        - It is advisable to confirm for yourself: check the md5sum against the fingerprint on the CRAN server
    - When prompted, I recommend that you change the install directory to root folder, e.g., `C:\R\R-4.3.2` . This will allow for installation of packages to the common library as opposed to a personal library.
        - I recommend this change because of how Windows assigns home folders. During initial setup Windows 10 prompted you to choose a username and whether you wanted your work stored locally or in your OneDrive folder. A worse case scenario? You select a user name with spaces, e.g.,"Mike Dohm," and you selected OneDrive. Both will cause challenges later for running and or installing packages for R.
        - I made a video for you. Video is about 26 minutes long; at 22 minute mark, video includes how to install R Commander (instructions provided Install R Commander).

https://youtu.be/upjmBieh3bM

2. **macOS PCs**, first you must download and install XQuartz from https://www.xquartz.org. Best to restart your mac after installing XQuartz then proceed to install R.
   After installing XQuartz, then return to https://cran.r-project.org, select Download for Mac(OS) X, and run the installer.
   Screenshots of the install process are available at the end of this page (click here or scroll down to **Macos setup, Screenshots**).
   - As of August 2021, be advised that there are two distinct R versions for your MacBook or iMac.
     - For MacBook or iMac with Apple's M1 or M2 ARM chip sets, download and install `R-4.2.2-arm64.pkg` .
       - If you recently purchased a new MacBook or iMac (2020 to present), then you probably have the M1 or M2 chipset (check by clicking the Apple icon, then selecting About this Mac or System Information (`/Applications/Utilities/System Information.app`)).
       - XQuartz version 2.8.5 works on macs with either the M1 or Intel chipsets.
     - For older MacBook or iMacs with Intel processors, download `R-4.2.2.pkg` .
     - ~~Depreciated 8/4/2021: Be advised that these instructions are for Intel-based macs. At the time of writing these instructions (April 2021), the installation of XQuartz and R should work on new M1-based macs. At the time of this writing (April 2021), however, R will not run natively on your M1 mac. It will run using Rosetta 2, an emulator that is included with your M1 mac. The R folks are busy working on a version that will run natively, which may be ready within a few months.~~
3. Don't forget to drag the When you first try to run R, you may get a popup screen which provides no option to start the app, and perhaps even a rather ominous option to move the app to Trash. Just close the warning message and right-click on the R app. A new screen pops up, which looks very much like the previous warning, but now you will see and option to open the app. Click on open to start R.
4. Like the message to Windows PC users, bypassing Apple's Gatekeeper to run R in no way will harm your computer — provided you have downloaded from official sites — R is a verified program. Apple has taken an aggressive line on developers and favors apps that are part of their app store.
5. LINUX distros. If your PC platform is Linux, then you should be comfortable with installation and updating of software. R base is already included in Debian distributions (e.g., Mint, Ubuntu). See https://cloud.r-project.org/ for additional instructions.
   - For Chromebook users, if you can install a Linux subsystem, then you can also install and run R. For instructions to install R see Levi's excellent writeup at levente.littvay.hu/chromebook/.

Note:

To install up-to-date R and RStudio, your Chromebook needs to have Intel or AMD CPU; my ASUS Chromebook has an ARM64 processor (MediaTek mt8183), and Levi's instructions don't apply. As of January 2024 I am pushing the installation process a bit on my little Chromebook and have successfully created the Linux container (Debian 11, Bullseye) and installed

base (and development) R version (4.0.4) included with the Linux distribution. In the next month I'll update progress with installing an R environment on ARM64-based Chromebook.

Test R

For both macOS and Windows PCs, successful installation of R on your computer installs base R programming language and a simple graphical user interface. Test your install by running code in the terminal (one line at a time) or via script:

Windows:

1. Rgui.exe (Windows PC)
2. File → New script

Enter code in script editor, e.g.,

```
myX <- c(1,2,3,4)
myY <- c(5,10,15,20)
plot(myY,myX)
```

3. Run code: Ctrl+R

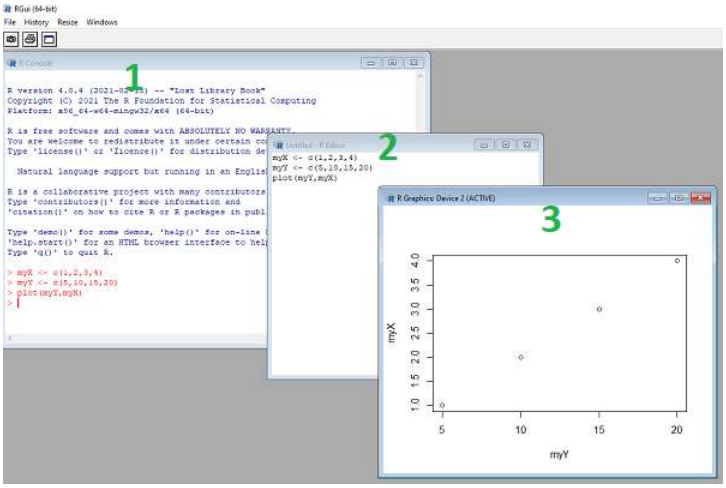


Figure $A.6.1$: Screenshot of RGui.exe (1), script editor (2), and results of `plot()` (3) on WinPC.

Mac:

1. R.app (macOS): run code in the terminal or via script
2. File → New Document

Enter code in script editor, e.g.,

```
myX <- c(1,2,3,4)
myY <- c(5,10,15,20)
plot(myY,myX)
```

3. Run code: Cmd+Enter

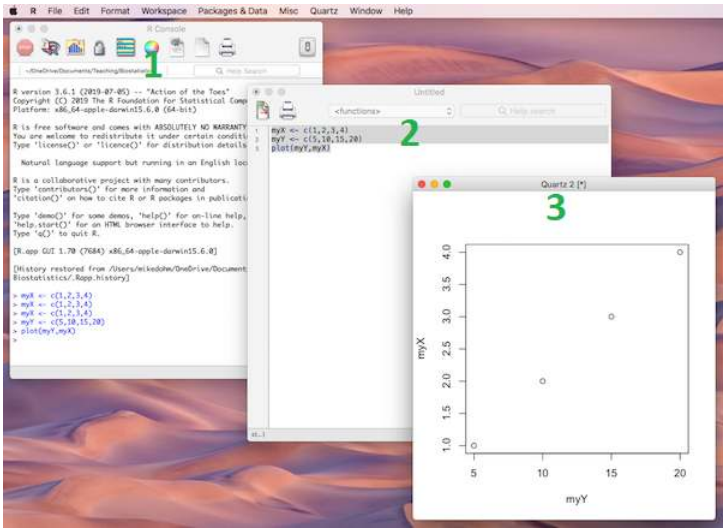


Figure $A.6.2$: Screenshot of R.app (1), script editor (2), and results of `plot()` (3) on macOS.

Many of you would like a video. Do a little search and you'll find plenty, although most are also showing how to install RStudio in addition to base R.

> **Note:**
>
> For my Biostatistics class, BI311, we typically will run R and use R Commander for scripting, without RStudio.

For BI311, we also use R Commander

R Commander is a package that adds function to R; it provides a familiar point-and-click interface to R, which allows the user to access functions via a drop-down menu system (Fox 2017).

Go to Install R Commander guide.

Run R in the "Cloud"

If you do not wish to install R, or, if you have a Chromebook and, therefore cannot gracefully install R, then there are alternatives; Run R in the Cloud. I'll list three ways to run R in the cloud for free. Go to Use R in the Cloud guide.

MacOS setup, Screenshots

Download R install package from R-project.org, then select the R install package from your Download folder.

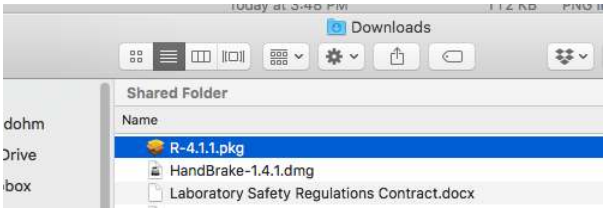


Figure $A.6.3$: R install package in MacOS Downloads.

First screen, R install for macOS. Select "continue".

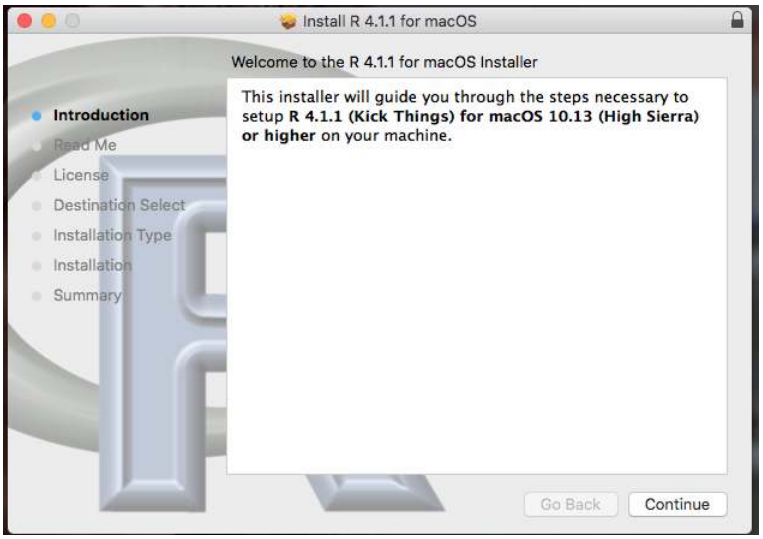


Figure $A.6.4$: Welcome screen for MacOS R installer.

Second screen, R install on macOS. Select “continue”.

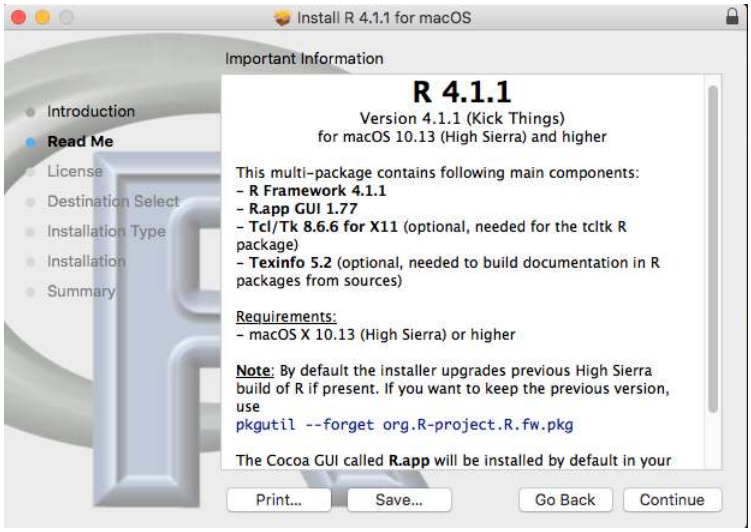


Figure $A.6.5$: Second screen for MacOS R installer.

Third screen, R install on macOS. Select “continue”.

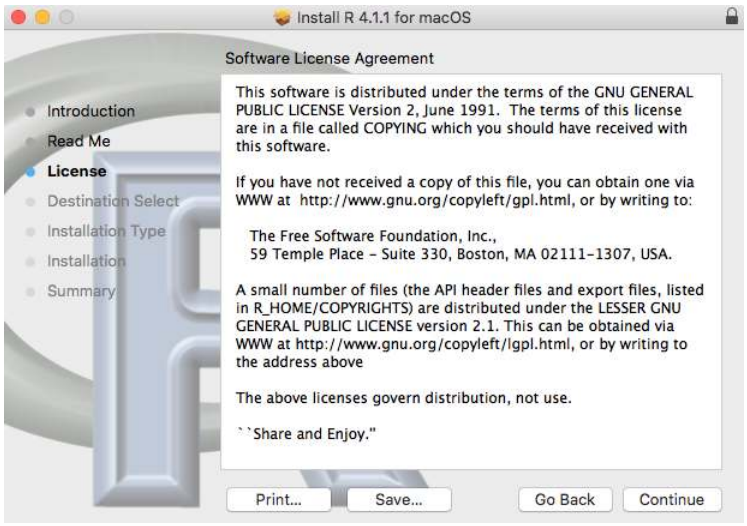


Figure $A.6.6$: Third screen for MacOS R installer.

Fourth screen, R install on macOS. Select “agree” to continue.

Figure $A.6.7$: Fourth screen for MacOS R installer, agree to terms.

Fifth screen, R install on macOS. Select “Install”.

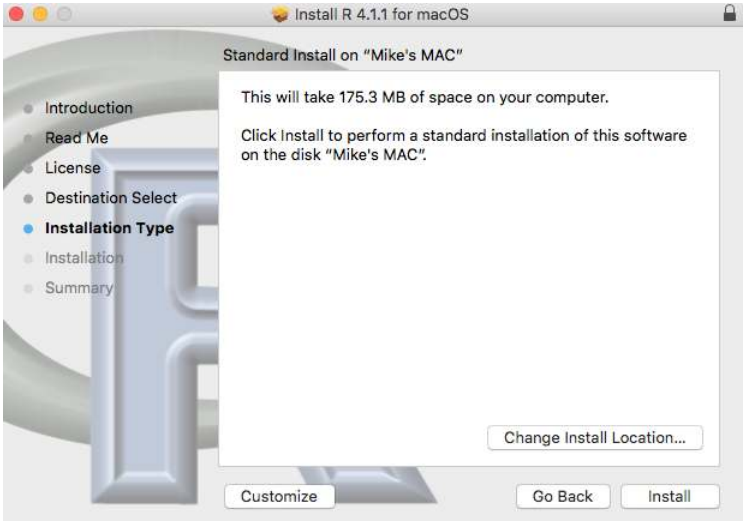


Figure $A.6.8$: Fifth screen for MacOS R installer, set install location and type.

Sixth screen, R install on macOS. Enter your username and password for your computer, then select “Install Software”.

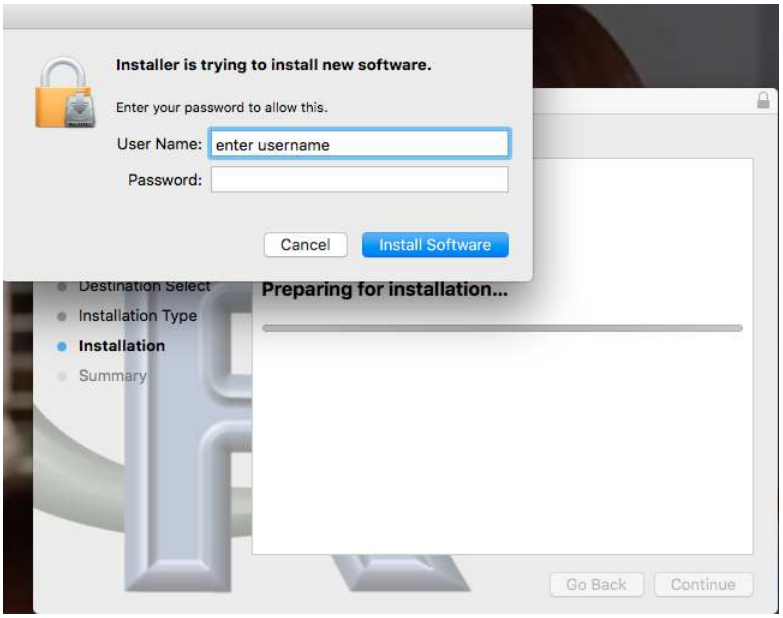


Figure $A.6.9$: Sixth screen for MacOS R installer, permission to install.

Seventh screen, R install on macOS. Several screens will pop up, reporting progress.

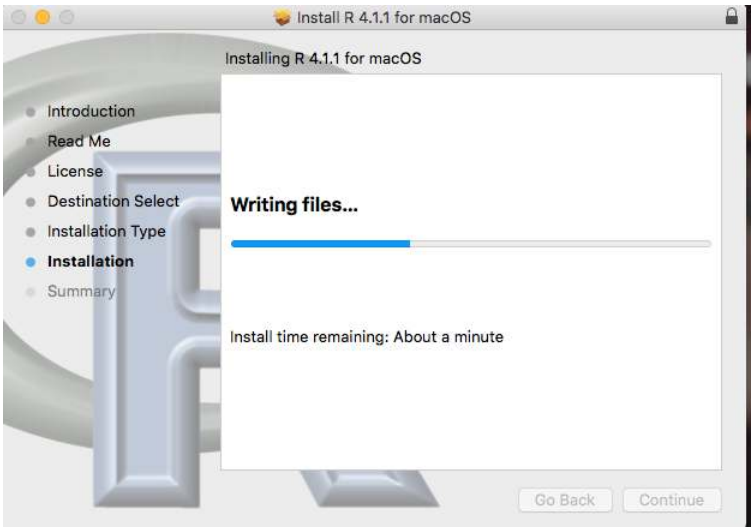


Figure $A.6.10$: Seventh screen for MacOS R installer, installation progress bar.

Eighth and final screen, R install on macOS. Select “Close”.

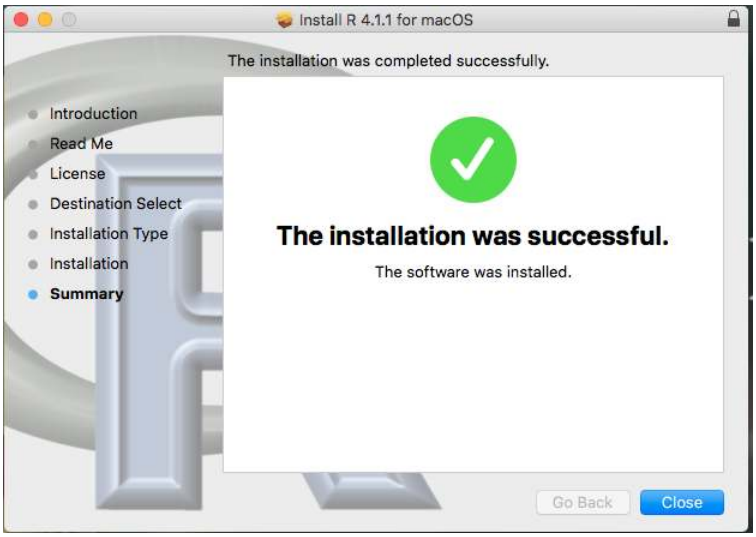


Figure $A.6.11$: Eighth and final screen for MacOS R installer, installation successful.

Optional — Keep or discard the install file. I keep and then do manual delete after I’ve confirmed the installation.

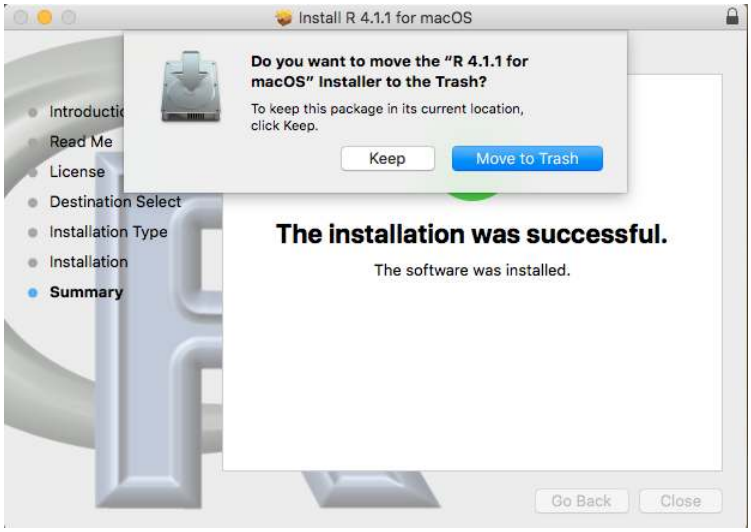


Figure $A.6.12$: Option to trash or keep the install file.

From Applications folder, start r.app. You should see the R Console.

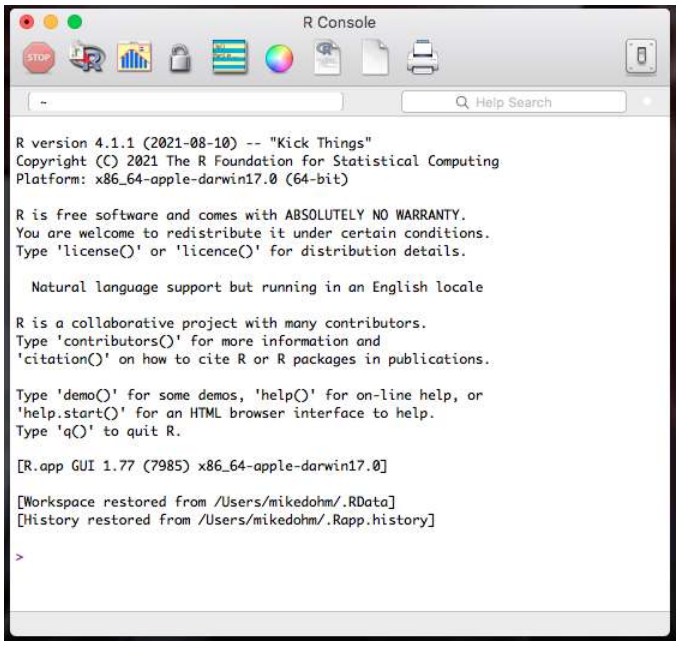


Figure $A.6.13$: R Console.

Wind10 setup, screenshots

Download from R-project.org, then right-click the R install package from your Download folder. Run as administrator.

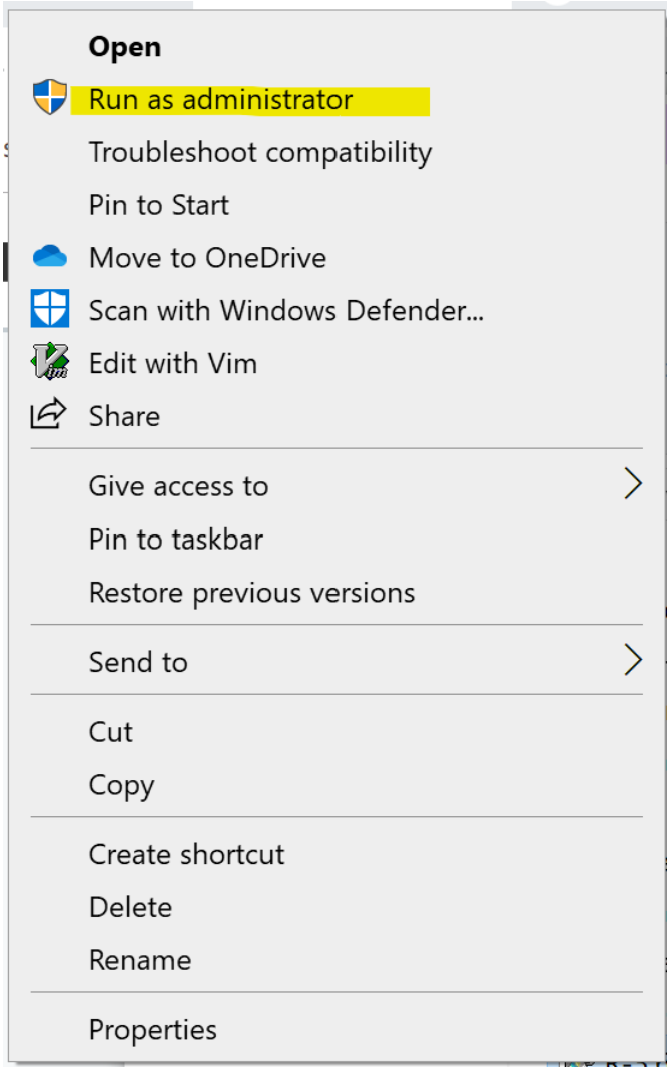


Figure $A.6.14$: Run R installer as administrator.

First screen, select language. Select OK to continue.

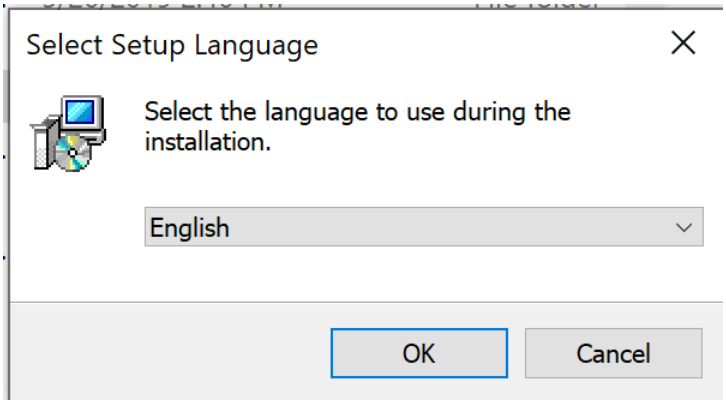


Figure $A.6.15$: Select language to use during installation.

Second screen, click Next to continue.

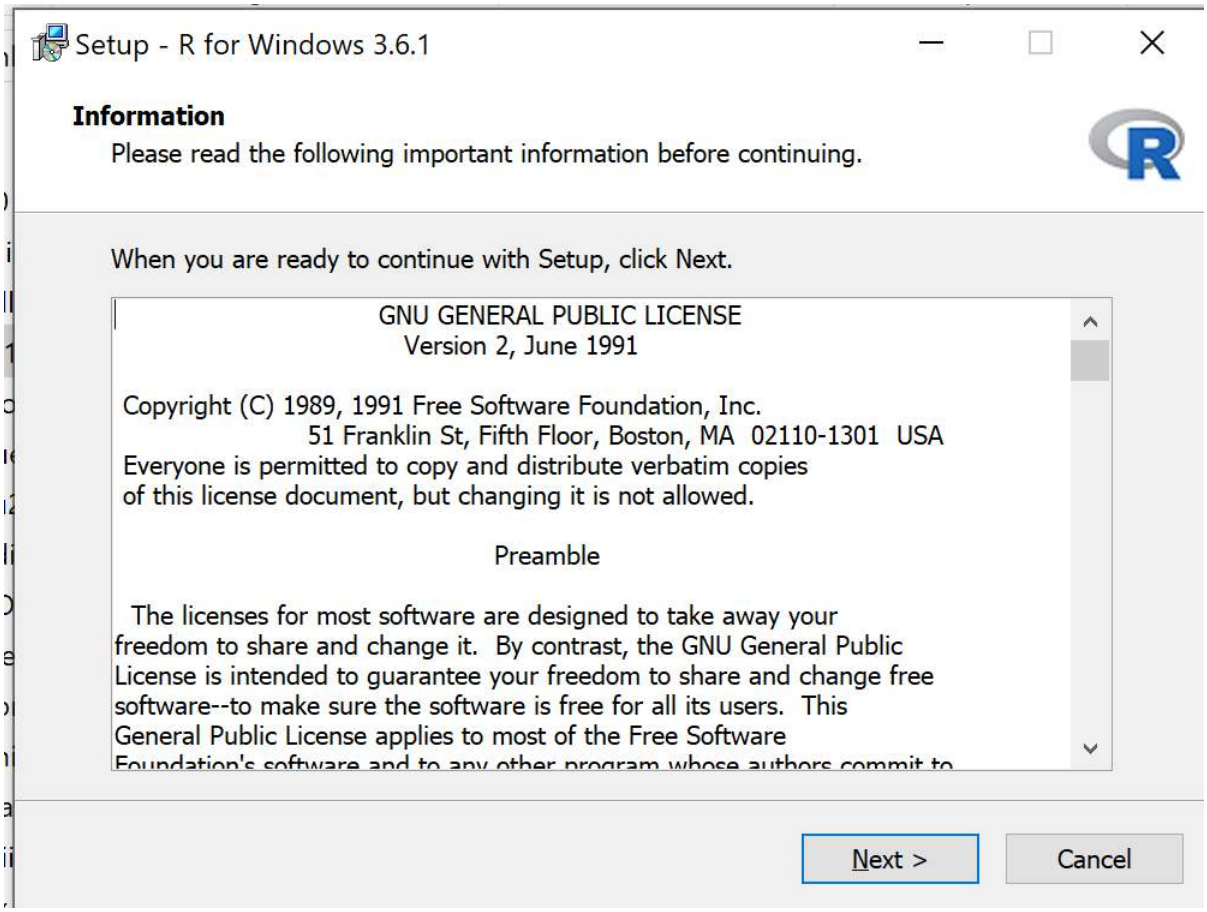


Figure $A.6.16$: Software information.

Third screen. Change the default location (show in the screenshot) to root folder, e.g., C:\R\R-4.1.1 (current version)

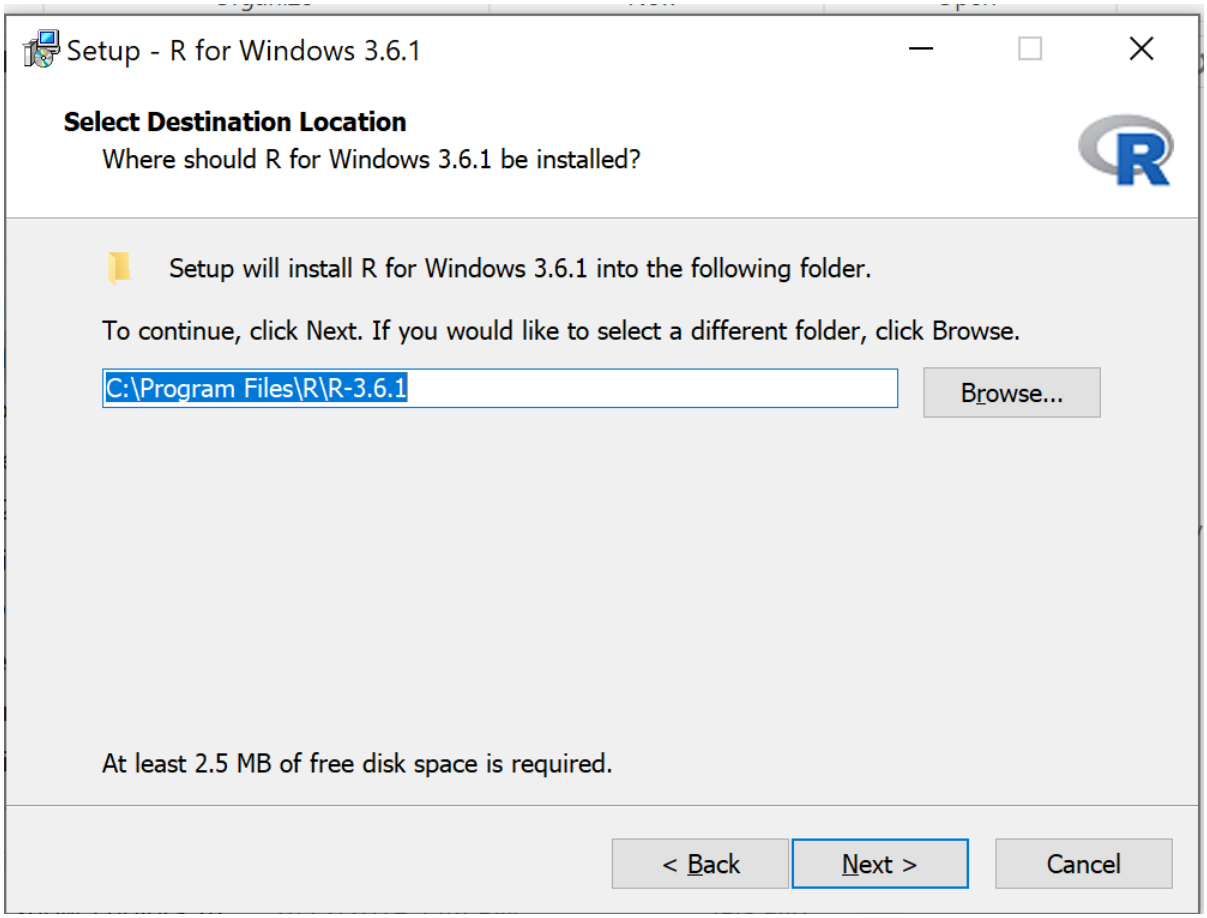


Figure $A.6.17$: Set location for R installation.

Fourth screen. Change startup options. Select Yes (customized startup) to continue.

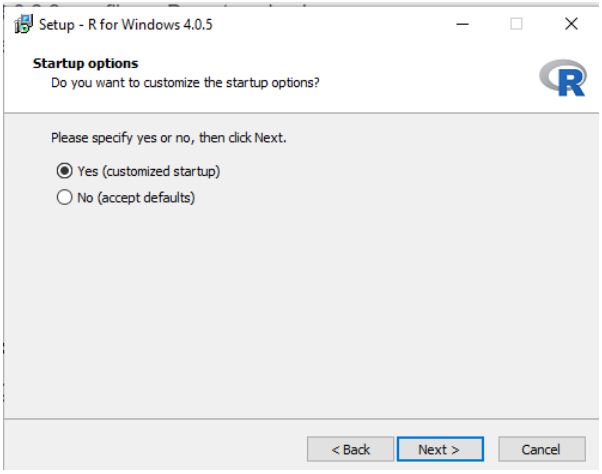


Figure $A.6.18$: Select customized startup.

Fifth screen, select SDI, then Next to continue.

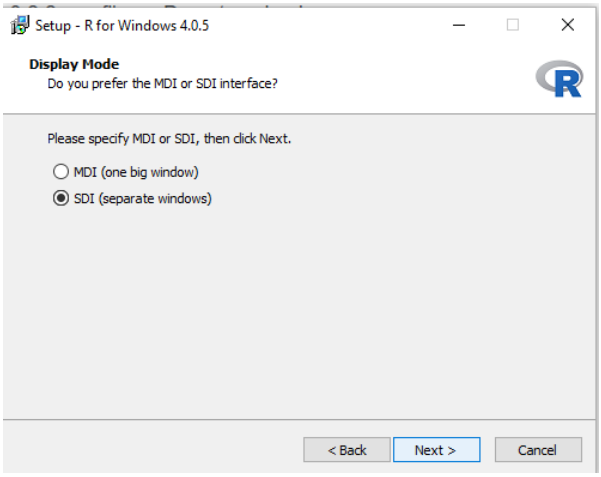


Figure $A.6.19$: Select the SDI display mode.

Sixth screen, select HTML help, then Next to continue

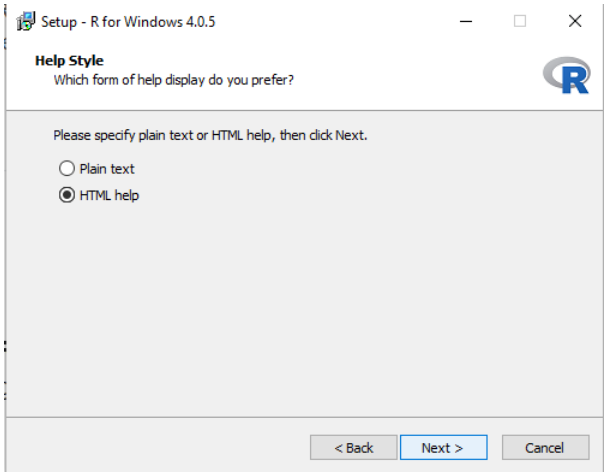


Figure $A.6.20$: Select the HTML help style.

Seventh screen, leave start menu folder as is (R), then Next to continue.

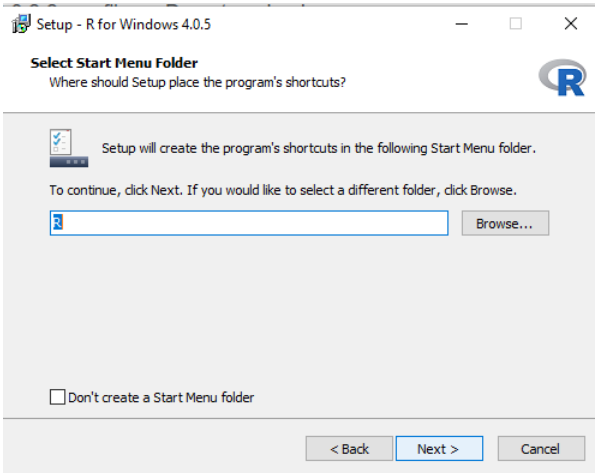


Figure $A.6.21$: Set the start menu folder for program shortcuts.

Eighth screen, check all boxes, then Next to continue.

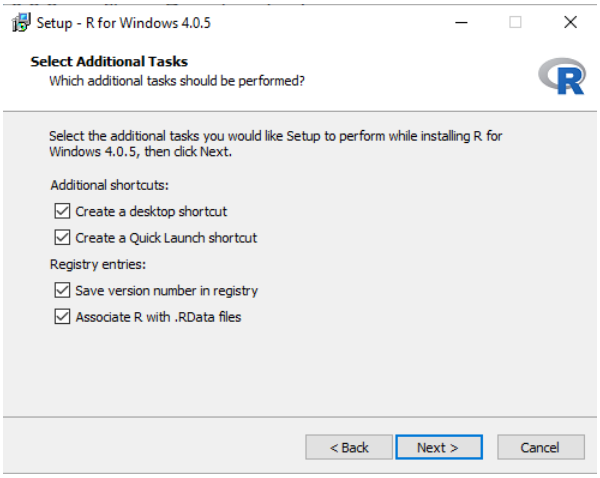


Figure $A.6.22$: Perform all additional tasks during installation.

Ninth screen, a series of status updates during the installation.

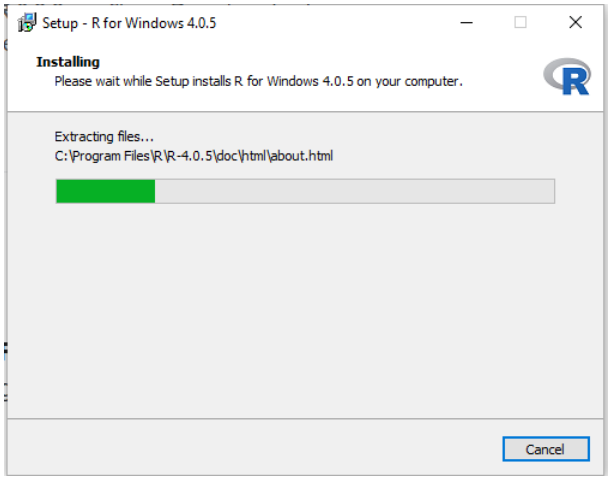


Figure $A.6.23$: Progress of installation.

Final screen, successful install.

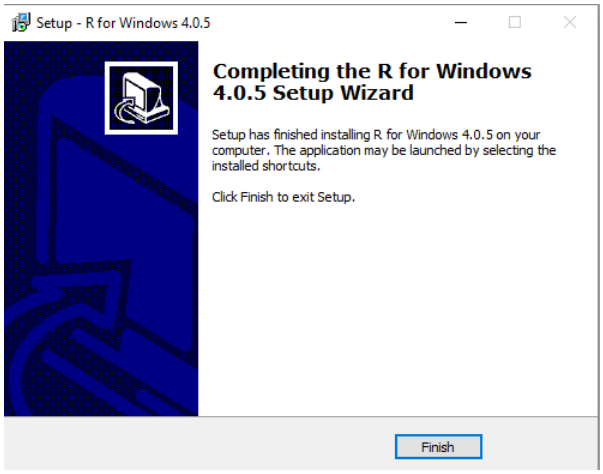


Figure $A.6.24$: R successfully installed on Windows.

A.7: Install R Commander

A quick guide about how to install R Commander onto your computer (LDE). You must have R installed and working correctly before proceeding to install the R Commander package. Click here to get the Install R guide.

> **Note:**
>
> If you plan to run R in the Cloud, you cannot install the R Commander package, which must be part of a local development environment.

Installation guides quickly become outdated. This page was last updated 15 August 2021 and describes working installation protocols at that time.

For BI311, we also use R Commander

R Commander is a package that adds function to R; it provides a familiar point-and-click interface to R, which allows the user to access functions via a drop-down menu system (Fox 2017). Thus, instead of writing code to run a statistical test, `Rcmdr` provides a simple menu driven approach to help students select and apply the correct statistical test. R Commander also provides access to `Rmarkdown` and a menu approach to rendering reports.

To install R Commander, enter the following code at the R prompt.

```
install.packages("Rcmdr")
```

In addition, download and install the following plugin:

```
install.packages("RcmdrMisc")
```

Note: You can combine requests as follows:

```
install.packages("Rcmdr", "RcmdrMisc", dependencies=TRUE)
```

Adding " `dependencies=TRUE` " will also install other packages that `Rcmdr` needs (which would get downloaded once you start Rcmdr for the first time).

If you have not set a mirror site, you'll be prompted to do so before you can download and install packages. I recommend 0-Cloud as default mirror site. Be advised: because our university shares a single public IP address, you may experience download delays if we all try to use the same mirror site at the same time.

To start R Commander, load the packages via the `library()` command.

```
library(Rcmdr)
```

Follow installation prompts. You can skip adding the " `otools` ," for now. However, `Rcmdr` will prompt you to install `otools` every time you start, so go ahead and install them at your convenience.

Mac users: To improve `Rcmdr` performance you must turn off "app nap." From `Rcmdr` , go to Tools, then select "Manage Mac OS X app nap for R.app …" Once you select "off" (click OK to apply), restart `Rcmdr` , the delay will be removed. Windows 10 folks don't have to contend with nap.

Add pandoc and LaTex support

To complete your R Commander installation you'll want to add additional document handling software support by adding LaTex and pandoc. In `Rcmdr` , select Tools, then Install Auxillary Software. Click OK, which will open links in your default browser to download pages for LaTex and pandoc. Download the files, follow the installation instructions for pandoc and LaTeX, then restart R and Rcmdr.

Here are direct links to the files, plus installation notes:

LaTeX

- MikTeX from https://miktex.org/download for Windows systems
- MacTeX from https://www.tug.org/mactex/ for MacOSX
    - Note the full installation is 4Gb. This is definitely overkill, but does provide all of the tools you could ever need. Other alternatives with smaller downloads require knowledge about what components are needed. So, in short, go for the large download, it's a simpler choice.
    - Be also advised: If you are running macOS before 10.14, you will have to jump through additional hoops to get macTeX installed.

pandoc

Windows 10/11

https://github.com/jgm/pandoc/releases/download/2.14.1/pandoc-2.14.1-windows-x86_64.msi

MacOS

https://github.com/jgm/pandoc/releases/download/2.14.2/pandoc-2.14.2-macOS.pkg

Test Rcmdr

Figure $A.7.1$ shows a basic R Commander session. Enter code in the script window (1), click on the Submit button to run the code, and results show up in the output window (2).

Copy and Paste
Image here.
Delete this
placeholder image

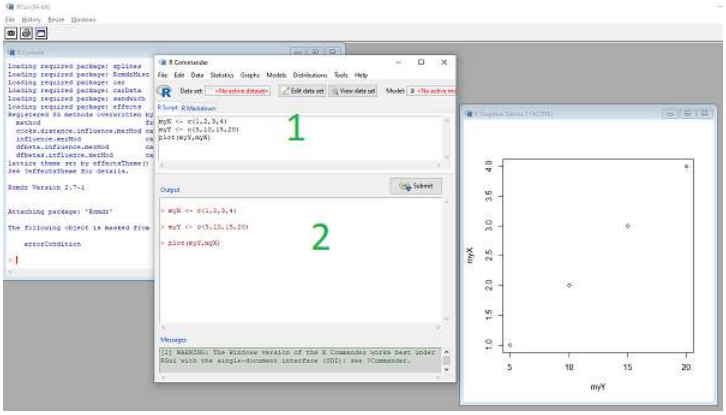


Figure $A.7.1$: Screenshot of basic R Commander session on WinPC.

Click on R Markdown tab, edit (e.g., replace with your own title and name), then click on the Generate Report button to create a pdf of your work, default file name is RcmdrMarkdown.pdf (Fig. $A.7.2$). If you do not have pandoc and LaTeX properly installed, then only an HTML document will be available as an option.

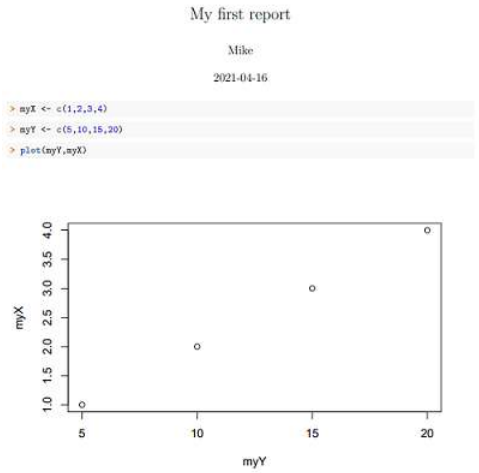


Figure $A.7.2$: Screenshot of a portion of RcmdrMarkdown.pdf.

A.8: Use R in the cloud

A quick guide about how to how to run R via a **cloud computing** (serverless) option.

Installation guides quickly become outdated. This page was last updated 19 August 2024 and describes working installation protocols at that time.

Quick links

- myCompiler
- CoCalc by SageMATH
- Google CoLaboratory
- RStudio in the cloud
- rdrr.io

Run R on your computer (i.e., local development environment or LDE)

This guide is about running R in the cloud, serverless options. For installing R and R Commander onto your own computer, see Install R and Install R Commander.

Run R "in the Cloud"

If you do not wish to install R, or if you have a ARM-based Chromebook and therefore cannot gracefully install R, then there are alternatives: Run R in the Cloud. I'll list five ways to run R in the cloud — run R on a server, not your own computer — for free.

Note: None of these options can run R Commander, which requires a local (on your computer) installation of R.

If you have a Chromebook, or you want to run R on your tablet (iPad, Kindle, etc.), you can't install R to any of these devices. However, you can access R via a serverless Cloud solution.

1. Run R code at Online R Compiler using **myCompiler**'s online IDE, link at https://www.mycompiler.io/online-r-compiler.

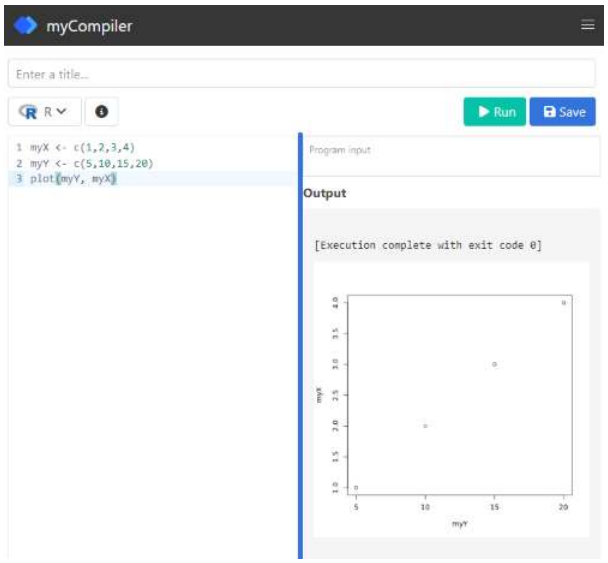


Figure $A.8.1$: Screenshot of myCompiler session.

2. Run code snippets in **CoCalc** by folks at SageMath and available at https://cocalc.com/. CoCalc uses Jupyter Notebooks, a wonderful, open-source project which supports interactive computer coding for many languages, including R and Markdown.

While CoLab is my go to, CoCalc is a really good student option — hint: I have my Systems Biology students use this option — includes SageMATH, python, GNU Octave and other software.

Create a free account (you'll then be able to save your code), or simply click "Run CoCalc Now" and check the box to agree to the terms to begin a session (Fig. $A.8.2$). Choose to open a new **Jupyter notebook**, then select R (system wide) from the choice of kernels.

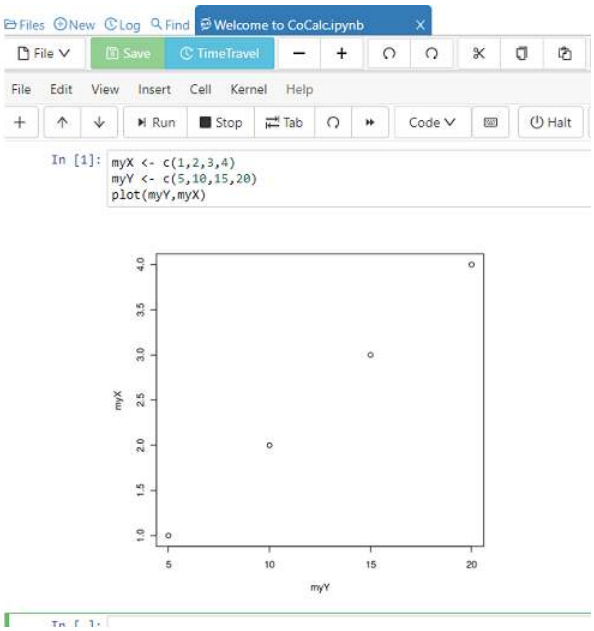


Figure $A.8.2$: Screenshot of CoCalc session.

You can load files from your computer for use in CoCalc sessions. There is also a version of the software you can download to your computer.

3. My favorite option, run R code snippets at Google **Colaboratory** (Fig. $A.8.3$). Like CoCalc, Colaboratory uses Jupyter Notebooks. Log into your Google account, then click https://colab.research.google.com/notebook#create=true&language=r , or try the tinyURL https://colab.to/r

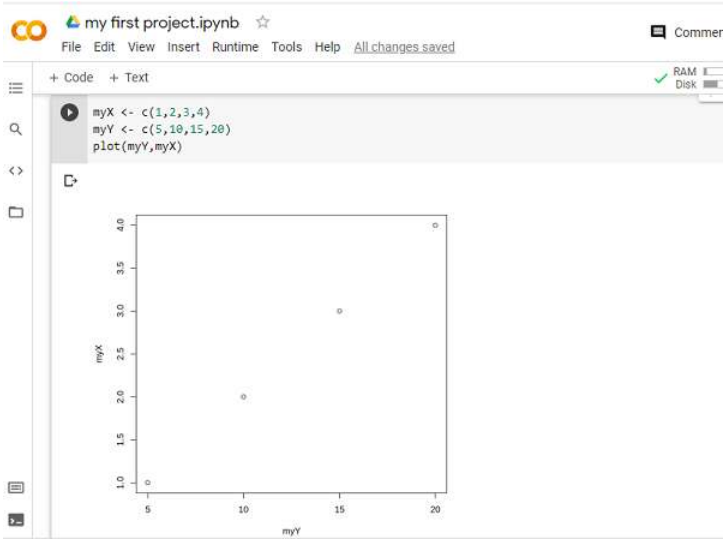


Figure $A.8.3$: Copy and Paste Caption here. (Copyright; author via source)

Colabs is worth the effort — you end up with a system to run R in your browser, it's free to use, and you can store/retrieve files from your Google Drive. This is my choice for Cloud computing, and it's the most generic solution. For more information, see post by Ed Adityawarman, How to use R in Google Colab . Colab Jupyter notebooks use Python by default. To run R, either use the link listed above each time you want to create a new R notebook, or add the following code snippet to your new notebook page

```
# activate R magic - must begin each R code with %%R
%load_ext rpy2.ipython
```

For all subsequent R code, start the section with

```
%%R
```

Note: You can install Jupyter onto your computer via Miniconda — **conda** is an open source package management system but then, you still would have to install R to your computer.

One real advantage of choosing CoLab, there are apps to run Google Colaboratory and Jupyter Notebooks on iPad/iPhone and for Android phones are available at Apple App Store and Google Play, respectively.

4. You can run **RStudio** at Posit Cloud. Registration and use is free for students. This works OK, but can be slow and it's hard to work on your own data. It does have the advantage of providing the familiar RStudio interface. Choose the free plan; Instructions to get started are at https://posit.cloud/plans. A screenshot of an RStudio cloud session is shown in Figure $A.8.4$.

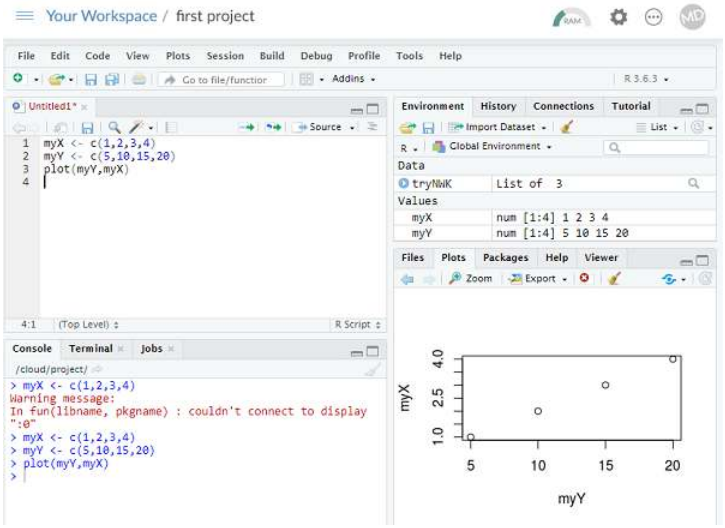


Figure $A.8.4$: Screenshot of RStudio Cloud session.

5. For limited use, i.e., you just need to run a little code to solve an assignment problem, you can run R **code snippets** in your browser at https://rdrr.io/snippets/ . You'll see many of my code embedded in this service so that you can run code snippets from my Chaminade University CANVAS pages.

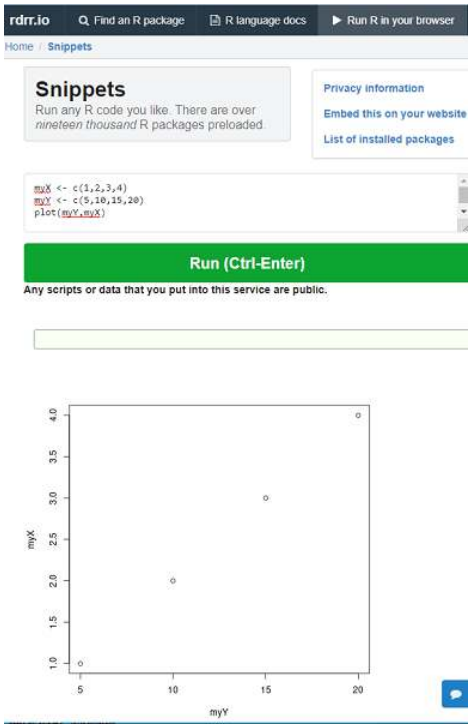


Figure $A.8.5$: Screenshot of rdrr.io/snippets session.

A.9: Jupyter notebook

Draft

Jupyter notebook, python. A “web-based computational environment”

Project homepage: https://jupyter.org/

Wikipedia

Besides the python kernel, Jupyter kernels include

Cytoscape

SageMATH

and, of course R, which along with python and Julia, is one of the core programming languages available in Jupyter. We present how to install the IRkernel on this page.

In the cloud

Access to Jupyter notebook was discussed for running R in the cloud.

Local installation

install latest python 3.12.4
https://www.python.org/

https://www.python.org/downloads/windows/
macOS universal installer
https://www.python.org/downloads/macos/

default python on macOS
see how to bash alias at https://stackoverflow.com/questions/...to-3-x-on-os-x

Open terminal
python3 –version
python3 -m pip –version
pip3 install jupyterlab

pip install jupyterlab
jupyter lab
browser opens http://localhost:8888/lab

Install IRkernel from CRAN

# Run R in terminal as administrator
sudo R
# At R prompt enter
install.packages(“IRkernel”)
# Making the kernel available to Jupyter
IRkernel::installspec(user = FALSE)

Run R as Jupyter Notebook

In the terminal, type at the bash shell line

```
jupyter lab
```

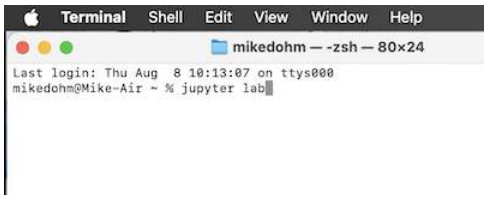


Figure $A.9.1$: Screenshot of terminal with Jupyter lab command.

Set working drive, then load kernel. Select the R kernel and create a new Notebook, Figure $A.9.2$ (i.e., don't select a Console).

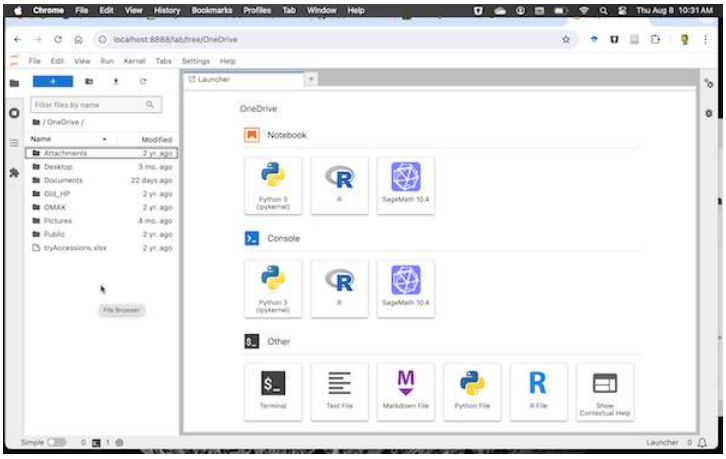


Figure $A.9.2$: Screenshot of Jupyter Lab launcher. Select R icon under Notebook to set IRkernel.

You should be ready to go.

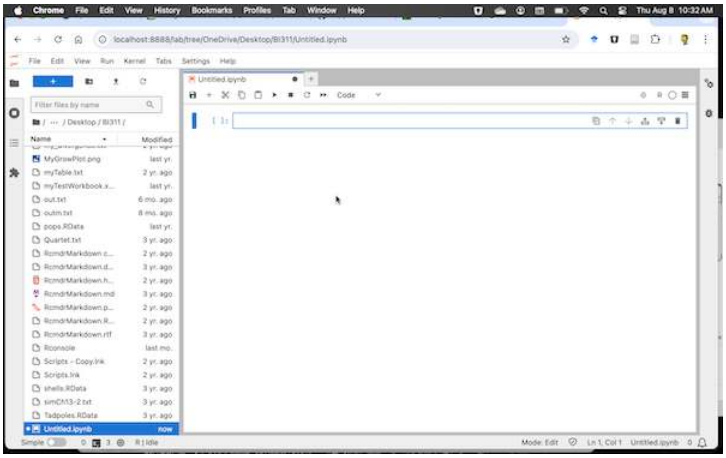


Figure $A.9.3$: Screenshot of Jupyter Notebook running the IRkernel.

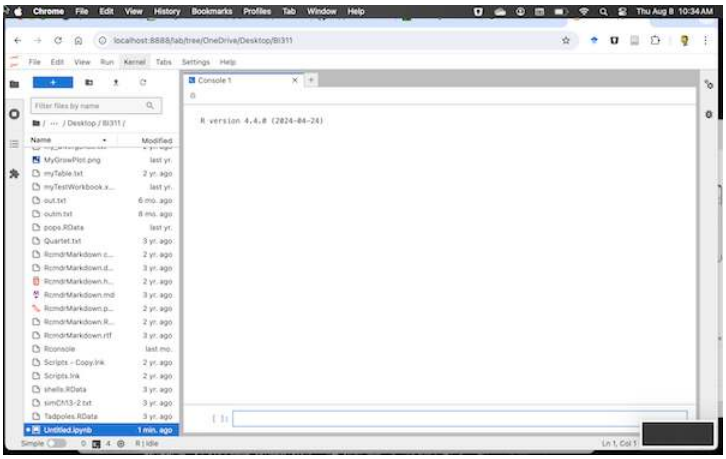


Figure $A.9.4$: Screenshot of Jupyter Console running the IRkernel.

It's easy to switch kernels. Let's say you started Jupyter Lab and notice that Python is running (Fig. $A.9.5$). Click on the kernel name — see green arrow in Figure $A.9.5$ — to bring up a popup menu, Fig $A.9.6$.

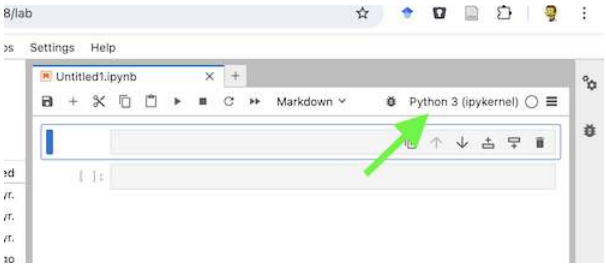


Figure $A.9.5$: Screenshot of Notebook with Python set as kernel.

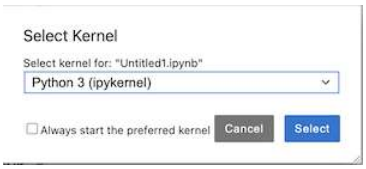


Figure $A.9.6$: Screenshot of select kernel popup menu.

Click on the drop arrow and select R kernel (figure $A.9.7$), then click on blue Select button.

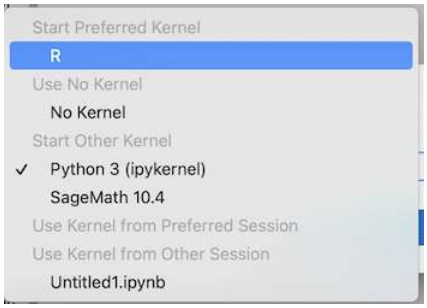


Figure $A.9.7$: Screenshot of installed kernels.

Once you select the R kernel from the dropdown, the kernel should be been successfully switched, as shown in Figure $A.9.8$.

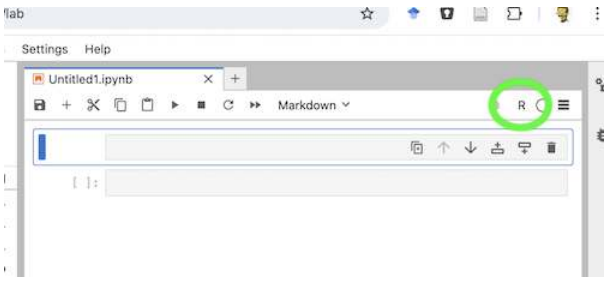


Figure $A.9.8$: Kernel successfully switched to R.

A.10: R packages

This page describes basic steps for **package installation** from a **CRAN mirror site** and how to update installed packages following installation of a new version of R. See at the end of this page for a list of packages described in Mike's Biostatistics Book.

Adding packages to base R installation

Installing R packages is straightforward, assuming the package is part of CRAN. Select a CRAN mirror site, e.g., **0-Cloud**, RStudio's mirror site.

```
chooseCRANmirror()
```

To find out what CRAN mirror was set for the current session use

```
findCRANmirror()
```

A list of mirror sites is stored on your computer once R is installed, see `CRAN_mirrors.csv` in the doc folder, e.g., `~/R-4.3.1/doc` .

Once the CRAN mirror is selected, and assuming you have the name of the package, e.g., `package.name` , then

```
install.packages("package.name")
```

will work.

Useful additional command options include

```
install.packages("package.name", dependencies=TRUE)
```

which will also download and install any additional packages required, and

```
install.packages("package.name", quiet=TRUE)
```

cuts down on the amount of screen output during installation.

If you receive the following **warning message**,

```
Warning: package 'package.name' is not available (for R version 4.3.2)
```

it may be possible that the package has not yet become available, but first double-check for typos.

Another warning message may be that a **binary version** is available, but a more recent **source version** is available, prompted by the question, *Do you want to install from sources the package which needs compilation?* In most cases, the answer is no. R will install a previous binary version. In order to install from source, `RTools` must be installed.

Update R packages after installing new R version

After updating to new version of R you'll need to download and update the user installed packages again. If you are running RStudio, see instructions here. For Win11 users you can download and run a package called `installr` , for macOS users download and install `updateR` , which will assist you to update R packages.

I prefer to run a script, modified from R-Bloggers.com. This script works on any operating system, but updates only CRAN packages (e.g., not github or Bioconductor).

Before installing the new version of base R, start up your current R installation and set your working directory, `setwd()` . Enter the following script to gather and save all installed R packages. Select CRAN mirror when prompted.

```
tmp <- installed.packages()
installedpkgs <- as.vector(tmp[is.na(tmp[,"Priority"]), 1])
save(installedpkgs, file="installed_old.rda")
```

Shutdown R, then install and start the new version of R (see Install R for help).

In the new version of R, set your working directory as above. Enter the following script

```
load(file="installed_old.rda")
tmp <- installed.packages()
installedpkgs.new <- as.vector(tmp[is.na(tmp[,"Priority"]), 1])
missing <- setdiff(installedpkgs, installedpkgs.new)
install.packages(missing)
update.packages(ask=FALSE)
```

Should be good to go. You can remove old R version installation.

Note:

To check installed packages, just view the object `installedpkgs` created earlier.

R packages used in Mike's Biostatistics Book

list updated 12 August 2024

package	chapter
agRee	16.5 – Instrument reliability and validity
ape	20.11 - Plot a Newick tree
baseline	20.3 - Baseline correction
BiocManager	20.11 - Plot a Newick tree
Bioconductor	20.11 - Plot a Newick tree
BiodiversityR	5.6 - Sampling from Populations
boot	19.2 - Bootstrap sampling
bootstrap	19.1 - Jackknife sampling
BSDA	11.4 - Two-sample effect size
cairoDevice	13.3 - Test assumption of normality
car	4.3 - Box plots
carData	4.1 - Bar (column) charts
cholera	2.3 - A brief history of (bio)statistics
clipr	4 - How to report statistics
combinat	6.3 - Combinations and permutations
confintr	19.2 - Bootstrap sampling
contingencytables	9.6 - McNemar's test
correlation	16.6 - Similarity and Distance

package	chapter
cranlogs	2.2 - Why do we use R Software?
datasets	4.5 - Scatter plots
digitize	12.3 - Fixed effects, random effects, and ICC
drc	20.10 - Growth equations and dose response calculations
effectsize	12.5 – Effect size for ANOVA
effsize	11.4 - Two-sample effect size
epiR	5.4 - Clinical trials
epitools	7.4 – Epidemiology: Relative risk and absolute risk, explained
exact2x2	9.6 – McNemar's test
factoextra	20.6 – Dimensional analysis
findpeaks	20.2 - Peak detection
forecast	20.5 - Time series
geepack	20.1 - Area under the curve
geeM	20.1 - Area under the curve
geodist	16.6 - Similarity and Distance
ggplot2	4.1 - Bar (column) charts
ggtree	20.11 - Plot a Newick tree
gplots	4.1 - Bar (column) charts
gtools	6.3 - Combinations and permutations
GrapheR	4.10 - Graph software
HH	12.4 - ANOVA from "sufficient statistics"
HistData	3.2 - Measures of Central Tendency
lattice	4.10 - Graph software
lmboot	19.1 - Jackknife sampling
irr	12.3 - Fixed effects, random effects, and ICC
MASS	12.4 - ANOVA from "sufficient statistics"
Matrix	20.1 - Area under the curve
mcp	12.6 - ANOVA post-hoc tests
MESS	20.1 - Area under the curve
mlr3misc	8.2 – The controversy over proper hypothesis testing
modeest	3.2 - Measures of Central Tendency
multcomp	12.6 - ANOVA posthoc tests
NCStats	3.3 - Measures of dispersion
nlopt	20.10 - Growth equations and dose response calculations

package	chapter
nortest	13.3 – Test assumption of normality
PairedData	10.3 – Paired t-test
peakDetection	20.2 - Peak detection
Phylotools	20.11 - Plot a Newick tree
Phytools	20.11 - Plot a Newick tree
plotly	4.10 - Graph software
plyr	4.1 - Bar (column) charts
polychor	16.4 – Spearman and other correlations
propCIs	7.6 - Confidence intervals
psa	20.6 – Dimensional analysis
psy	12.3 - Fixed effects, random effects, and ICC
psych	3.2 - Measures of Central Tendency
pwr	11.5 - Power analysis in R
random	6.6 - Continuous distributions
rattle	13.3 - Test assumption of normality
Rcmdr	1.1 - A quick look at R and R Commander
RcmdrMisc	1.1 - A quick look at R and R Commander
RcmdrPlugin.EBM	4.4 - Mosaic plots
RcmdrPlugin.EZR	11.5 - Power analysis in R
RcmdrPlugin.HH	12.4 - ANOVA from "sufficient statistics"
RcmdrPlugin.KMggplot2	4.1 - Bar (column) charts
RcmdrPlugin.mosaic	4.4 - Mosaic plots
RcmdrPlugin.survival	20.9 - Survival analysis
Rcolorbrewer	4.4 - Mosaic plots
reshape2	4.6 - Adding a second Y axis
rgl	18.1 - Multiple Linear Regression
Rmisc	3.5 - Statistics of error
ROCR	20.1 - Area under the curve
rptR	12.3 - Fixed effects, random effects, and ICC
RGtk2	13.3 - Test assumption of normality
season	20.5 – Time series
shotGroups	3.5 - Statistics of error
stats	4 – How to report statistics
survival	3.1 - Data types

package	chapter
tanggle	20.11 - Plot a Newick tree
Ternary	4.8 - Ternary plots
testequavar	13.4 - Tests for Equal Variances
tidyverse	4.3 - Box plot
tigerstats	8.4 - Tails of a test
timeseries	20.5 - Time series
TOSTER	16.1 - Product-moment correlation
vegan	20.8 - Diversity indexes
WRS2	3.3 - Measures of dispersion

A.11: List of R commands

List and links to R commands (followed with parentheses), R packages, and R Commander menu selections

Link to terms: Index Mike's Biostatistics Book

Click on name of command to take you to the chapter and section where the command is presented. Note that you may need to scroll down on the page to view the code and command.

R commands

.RProfile
aov()
ave()
c()
chisq.text()
data()
data.frame()
Dotplot()
dplyr()
epiR; Ch07.2
epi.conf()
exp()
geosd()
head()
help()
kruskal.test()
log()
mad()
mean()
median()
names()
pchisq()
plot()
pnorm()
pnormGC()
qchisq()
qt()
quantile()
range()
rank()
require()
RGUI menu: File → New script
round()
scan()
sd()
seq()
stack()
summary()
table()
t.test()
tapply()
with()

R Commander menu selections

Rcmdr: Distributions → Continuous distributions → Normal distribution → Normal quantiles…
Rcmdr: File → Exit
Rcmdr: Manage Mac OS X app nap for R.app…

A.12: Free apps for bioinformatics

Mike's recommended free and/or open-source apps for bioinformatics on macOS or Windows 11 PCs.

> **Note:**
>
> Chrome OS users: If your device is Intel-based, then it is possible to install many of the apps (or equivalents) listed via activating LINUX on your device. This route is only advisable if you are willing and able to do some pretty serious installation work.

Statistics

R Core Team (2024). **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Python programming language and libraries: panda, Numby, Scipy, Matplotlib, Biopython, PySB, and others

Mathematics software

GNU Octave, free and open source, uses similar programming syntax to MATLAB, https://octave.org/

SageMATH, python based, free and open source alternative to MATLAB and Wolfram Mathematica, https://www.sagemath.org/

Image, Drawing, Concept & Mind maps, Flow charts

Newt Editor, a free, web based, open source viewer and editor for biology pathways, https://newteditor.org/index.html

Draw.io, used to create work flows, mind maps, https://app.diagrams.net/

FreeCAD, open source "parametric 3D modeler", https://www.freecad.org/

GIMP, GNU Image Manipulation Program, at https://www.gimp.org/

ImageJ2, JAVA-based, https://imagej.net/software/imagej2/

Krita, free and open source painting program, https://krita.org/en/

Preview, macos only

Video editor

Shortcut, at https://shotcut.org/

Handbrake, at https://handbrake.fr/

Screen recording, Video streaming

OBS Studio, at https://obsproject.com/

QuickTimePlayer, macos only

Office suite (compatible with Microsoft Word)

LibreOffice, at https://www.libreoffice.org/

Reference manager

Zotero, https://www.zotero.org/

Digital Notebook

Jupyter Lab and Notebooks, python based, web-based computational notebooks. In addition to python, Jupyter supports Julia and R Programming languages as well as Cytoscape, GNU Octave, SageMATH and other software. https://docs.jupyter.org/en/latest/index.html

OneNote, at https://www.onenote.com/

Bioinformatics tools

Bioconductor, at https://www.bioconductor.org/

Unipro UGENE, at http://ugene.net/

Cytoscape, at https://cytoscape.org/

MEGA, at https://www.megasoftware.net/

Coding IDE and code editor

Jupyter Lab

Visual Studio with Python application, Community edition, https://visualstudio.microsoft.com/vs/features/python/

Posit team (2024). **RStudio**: Integrated Development Environment for R. Posit Software, PBC, Boston, MA URL http://www.posit.co/.

Index

Index to key terms in this eBook; 519 terms, 598 links (last updated 7 December 2023)

R commands used in this book available at List of R commands

Click on term to take you to chapter and section where the word is first presented; subsequent references are noted by chapter (e.g., Ch08.3 refers to Chapter 8.3).

Note you may need to scroll down on the page to view the word: use browser Find (Ctrl+F or Cmd+F).

Detailed Licensing

Overview

Title: Mike's Biostatistics Book (Dohm)

Webpages: 187

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 97.9% (183 pages)
- Undeclared: 2.1% (4 pages)

By Page