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.