BIOSTATISTICS - OPEN LEARNING TEXTBOOK

LibreTexts™

Biostatistics - Open Learning Textbook

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

Unit 4A: Introduction to Statistical Inference

Unit 4B: Inference for Relationships

Licensing

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

CHAPTER OVERVIEW

Preliminaries

In this introductory section for PHC 6050 and PHC 6052, we will:

- Define statistics and biostatistics
- List the five steps in a typical research project and discuss the roles biostatistics can play in each
- Introduce the Big Picture of Statistics, which is the foundation of our course, and define and discuss its four components
- Introduce fundamental definitions related to data, datasets, and variables.
- Explain the different types/classifications of variables and introduce why this is important in biostatistics

Here are links to a few other online materials similar to 6060/6052 which you may find useful as secondary references

- Materials on which we based the 6050/6052 course: https://oli.cmu.edu/jcourse/webui/guest/join.do?section=probstat (click ENTER COURSE or create an account and login)
- Very similar course through Penn State: https://onlinecourses.science.psu.edu/stat500/
- http://onlinestatbook.com/index.html
- http://www.seeingstatistics.com/
- http://www.jerrydallal.com/LHSP/LHSP.HTM
- https://www.openintro.org/stat/
- https://en.wikiversity.org/wiki/Introduction_to_Statistical_Analysis
- https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/index.htm

Role of Biostatistics

Our first course objective will be addressed throughout the semester in that you will be adding to your understanding of biostatistics in an ongoing manner during the course.

CO-1: Describe the roles biostatistics serves in the discipline of public health.

What is Biostatistics?

Learning Objectives

LO 1.1: Define statistics and biostatistics.

Biostatistics is the application of **statistics** to a variety of topics in biology. In this course, we tend to focus on biological topics in the health sciences as we learn about statistics.

In an introductory course such as ours, there is essentially no difference between "biostatistics" and "statistics" and thus you will notice that we focus on learning "statistics" in general but use as many examples from and applications to the health sciences as possible.

Note

Statistics is all about **converting data into useful information**. Statistics is therefore a process where we are:

- collecting data,
- summarizing data, and
- interpreting data.

The following video adapted from material available from Johns Hopkins – Introduction to Biostatistics provides a few examples of statistics in use.

Video

Statistics Examples (3:14)

The following reading from the online version of Little Handbook of Statistical Practice contains excellent comments about common reasons why many people feel that "statistics is hard" and how to overcome them! We will suggest returning to and reviewing this document as we cover some of the topics mentioned in the reading.

Reading

Is Statistics Hard? (≈ 1500 words)

Steps in a Research Project

Learning Objectives

LO 1.2: Identify the steps in a research project.

In practice, every **research project** or study involves the following **steps**.

1. Planning/design of study
2. Data collection
3. Data analysis
4. Presentation
5. Interpretation

The following video adapted from material available at Johns Hopkins – Introduction to Biostatistics provides an overview of the steps in a research project and the role biostatistics and biostatisticians play in each step.

Video

Role of Biostatistics in the Steps of a Research Project (5:23)

(Optional) Outside Reading: Role of Biostatistics in Modern Medicine (≈ 1000 words)

The Big Picture

CO-1: Describe the roles biostatistics serves in the discipline of public health.

Throughout the course, we will add to our understanding of the definitions, concepts, and processes which are introduced here. You are not expected to gain a full understanding of this process until much later in the course!

To really understand how this process works, we need to put it in a context. We will do that by introducing one of the central ideas of this course, the **Big Picture of Statistics**.

We will introduce the Big Picture by building it gradually and explaining each component.

At the end of the introductory explanation, once you have the full Big Picture in front of you, we will show it again using a concrete example.

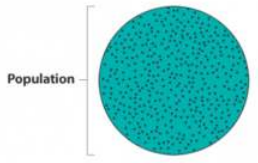
Learning Objectives

LO 1.3: Identify and differentiate between the components of the Big Picture of Statistics

Video

The Big Picture of Statistics (4:59)

The process of statistics starts when we identify what group we want to study or learn something about. We call this group the **population**.



Note that the word "population" here (and in the entire course) is not just used to refer to people; it is used in the more broad statistical sense, where population can refer not only to people, but also to animals, things etc. For example, we might be interested in:

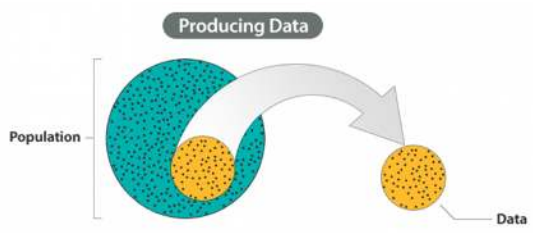
- the opinions of the population of U.S. adults about the death penalty; or
- how the population of mice react to a certain chemical; or
- the average price of the population of all one-bedroom apartments in a certain city.

Note

The **population**, then, is the entire group that is the target of our interest.

In most cases, the population is so large that as much as we might want to, there is absolutely no way that we can study all of it (imagine trying to get the opinions of all U.S. adults about the death penalty…).

A more practical approach would be to examine and collect data only from a sub-group of the population, which we call a **sample**. We call this first component, which involves choosing a sample and collecting data from it, **Producing Data**.



Note

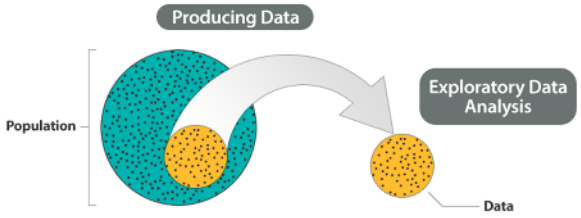
A **sample** is a s subset of the population from which we collect data.

It should be noted that since, for practical reasons, we need to compromise and examine only a sub-group of the population rather than the whole population, we should make an effort to choose a sample in such a way that it will represent the population well.

For example, if we choose a sample from the population of U.S. adults, and ask their opinions about a particular federal health care program, we do not want our sample to consist of only Republicans or only Democrats.

Once the data have been collected, what we have is a long list of answers to questions, or numbers, and in order to explore and make sense of the data, we need to summarize that list in a meaningful way.

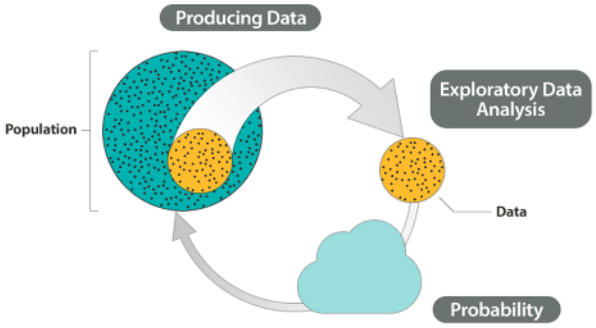
This second component, which consists of summarizing the collected data, is called **Exploratory Data Analysis** or **Descriptive Statistics**.



Now we've obtained the sample results and summarized them, but we are not done. Remember that our goal is to study the population, so what we want is to be able to draw conclusions about the population based on the sample results.

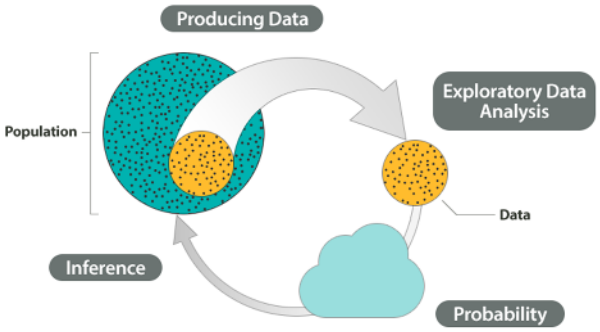
Before we can do so, we need to look at how the sample we're using may differ from the population as a whole, so that we can factor that into our analysis. To examine this difference, we use **Probability** which is the third component in the big picture.

The third component in the Big Picture of Statistics, **probability** is in essence the "machinery" that allows us to draw conclusions about the population based on the data collected in the sample.



Finally, we can use what we've discovered about our sample to draw conclusions about our population.

We call this final component in the process **Inference**.



This is the **Big Picture of Statistics**.

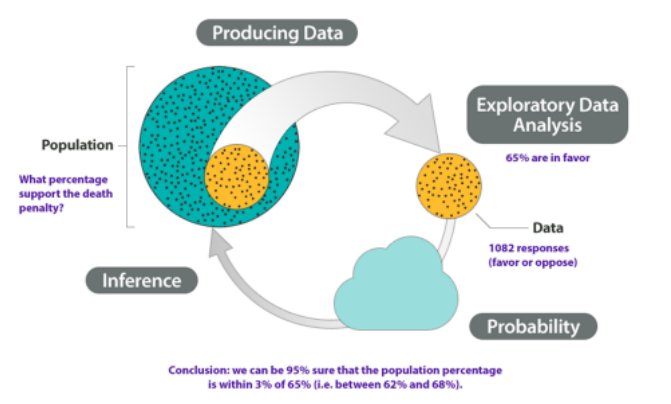
EXAMPLE: Polling Public Opinion

At the end of April 2005, a poll was conducted (by ABC News and the Washington Post), for the purpose of learning the opinions of U.S. adults about the death penalty.

1. Producing Data: A (representative) sample of 1,082 U.S. adults was chosen, and each adult was asked whether he or she favored or opposed the death penalty.

2. Exploratory Data Analysis (EDA): The collected data were summarized, and it was found that 65% of the sampled adults favor the death penalty for persons convicted of murder.

3 and 4. Probability and Inference: Based on the sample result (of 65% favoring the death penalty) and our knowledge of probability, it was concluded (with 95% confidence) that the percentage of those who favor the death penalty in the population is within 3% of what was obtained in the sample (i.e., between 62% and 68%). The following figure summarizes the example:



Course Structure

The structure of this entire course is based on the big picture.

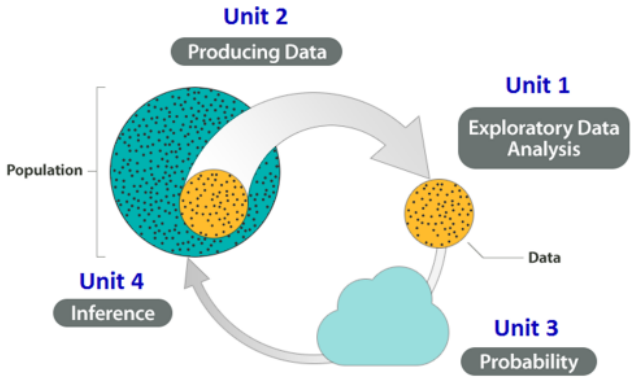
The course will have 4 units; one for each of the components in the big picture.

As the figure below shows, even though it is second in the process of statistics, we will start this course with exploratory data analysis (EDA), continue to discuss producing data, then go on to probability, so that at the end we will be able to discuss inference.

The main reasons we begin with EDA is that we need to understand enough about what we want to do with our data before we can discuss the issues related to how to collect it!!

This also allows us to introduce many important concepts early in the course so that you will have ample time to master them before we return to inference at the end of the course.

The following figure summarizes the structure of the course.



As you will see, the Big Picture is the basis upon which the entire course is built, both conceptually and structurally.

We will refer to it often, and having it in mind will help you as you go through the course.

Types of Variables

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

CO-7: Use statistical software to analyze public health data.

Classifying Types of Variables

Learning Objectives

LO 4.1: Determine the type (categorical or quantitative) of a given variable.

Learning Objectives

LO 4.2: Classify a given variable as nominal, ordinal, discrete, or continuous.

Video

Types of Variables (3 Parts; 13:25 total time)

Variables can be broadly classified into one of two **types**:

- Quantitative
- Categorical

Below we define these two main types of variables and provide further sub-classifications for each type.

Note

Categorical variables take **category** or **label** values, and place an individual into one of several **groups**.

Categorical variables are often further classified as either:

- **Nominal,** when there **is no natural ordering among the categories**.

  Common examples would be gender, eye color, or ethnicity.

- **Ordinal**, when there **is a natural order among the categories**, such as, ranking scales or letter grades.

  However, ordinal variables are still categorical and do not provide precise measurements.

  Differences are not precisely meaningful, for example, if one student scores an A and another a B on an assignment, we cannot say precisely the difference in their scores, only that an A is larger than a B.

Note

Quantitative variables take **numerical** values, and represent some kind of **measurement**.

Quantitative variables are often further classified as either:

- **Discrete**, when the variable takes on a **countable** number of values.

  Most often these variables indeed represent some kind of **count** such as the number of prescriptions an individual takes daily.

- **Continuous**, when the variable **can take on any value in some range of values**.

  Our precision in measuring these variables is often limited by our instruments.

  Units should be provided.

Common examples would be height (inches), weight (pounds), or time to recovery (days).

One special variable type occurs when a variable has only two possible values.

Note

A variable is said to be **Binary** or **Dichotomous**, when there are only two possible levels.

These variables can usually be phrased in a "yes/no" question. Whether nor not someone is a smoker is an example of a binary variable.

Currently we are primarily concerned with classifying variables as either categorical or quantitative.

Sometimes, however, we will need to consider further and sub-classify these variables as defined above.

These concepts will be discussed and reviewed as needed but here is a quick practice on sub-classifying categorical and quantitative variables.

Did I Get This?

Types of Variables

EXAMPLE: Medical Records

Let's revisit the dataset showing medical records for a sample of patients

	Variables					
Individuals	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race
Patient #1	M	59	175	69	1	White
Patient #2	F	67	140	62	2	Black
Patient #3	F	73	155	59	1	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	90	72	1	White

In our example of medical records, there are several variables of each type:

- Age, Weight, and Height are **quantitative** variables.
- Race, Gender, and Smoking are **categorical** variables.

Comments:

- Notice that the values of the **categorical** variable Smoking have been **coded** as the numbers 0 or 1.

It is quite common to code the values of a categorical variable as numbers, but you should remember that these are just codes.

They have no arithmetic meaning (i.e., it does not make sense to add, subtract, multiply, divide, or compare the magnitude of such values).

Usually, if such a coding is used, all categorical variables will be coded and we will tend to do this type of coding for datasets in this course.

- Sometimes, **quantitative** variables are **divided into groups** for analysis, in such a situation, although the original variable was quantitative, the variable analyzed is categorical.

A common example is to provide information about an individual's Body Mass Index by stating whether the individual is underweight, normal, overweight, or obese.

This categorized BMI is an example of an ordinal categorical variable.

- **Categorical** variables are sometimes called qualitative variables, but in this course we'll use the term "categorical."

Software Activity

Learning Objectives

LO 7.1: View a dataset in EXCEL, text editor, or other spreadsheet or statistical software.

Learning Objectives

LO 4.1: Determine the type (categorical or quantitative) of a given variable.

Learn By Doing:

Exploring a Dataset using Software

Why Does the Type of Variable Matter?

Note

The **types of variables** you are analyzing **directly relate to the available** descriptive and inferential **statistical methods**.

It is important to:

- **assess how you will measure the effect of interest** and
- **know how this determines the statistical methods you can use.**

As we proceed in this course, we will continually emphasize the **types of variables** that are **appropriate for each method we discuss**.

For example:

EXAMPLE:

To compare the number of polio cases in the two treatment arms of the Salk Polio vaccine trial, you could use

- Fisher's Exact Test
- Chi-Square Test

To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could use

- Two-sample t-Test
- Wilcoxon Rank-Sum Test

(Optional) Great Resource: : UCLA Institute for Digital Research and Education – What statistical analysis should I use?

What is Data?

CO-1: Describe the roles biostatistics serves in the discipline of public health.

Before we jump into Exploratory Data Analysis, and really appreciate its importance in the process of statistical analysis, let's take a step back for a minute and ask:

What do we really mean by **data**?

Learning Objectives

LO 1.4: Define basic terms regarding data and recognize common variations in terminology.

Video

What is Data? (2:49)

Data are pieces of information about **individuals** organized into **variables**.

- By an **individual**, we mean a particular person or object.
- By a **variable**, we mean a particular characteristic of the individual.

A **dataset** is a set of data identified with a particular experiment, scenario, or circumstance.

Datasets are typically displayed in tables, in which rows represent individuals and columns represent variables.

EXAMPLE: Medical Records

The following dataset shows medical records for a sample of patients.

	Variables					
Individuals	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race
Patient #1	M	59	175	69	1	White
Patient #2	F	67	140	62	2	Black
Patient #3	F	73	155	59	1	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	90	72	1	White

In this example,

- the **individuals** are patients,
- and the **variables** are Gender, Age, Weight, Height, Smoking, and Race.

Each **row**, then, gives us all of the information about a particular **individual** (in this case, patient), and each **column** gives us information about a particular **characteristic** of all of the patients.

Individuals, Observations, or Cases

Note

The rows in a dataset (representing **individuals**) might also be called **observations**, **cases**, or a description that is specific to the individuals and the scenario.

For example, if we were interested in studying flu vaccinations in school children across the U.S., we could collect data where each observation was a

- student

- school
- school district
- city
- county
- state

Each of these would result in a different way to investigate questions about flu vaccinations in school children.

Independent Observations

> **Note**
>
> In our course, we will present methods which can be used when the **observations** being analyzed are **independent of each other**. If the observations (rows in our dataset) are not independent, a more complex analysis is needed.Clear violations of independent observations occur when
>
> - we have more than one row for a given individual such as if we gather the same measurements at many different times for individuals in our study
> - individuals are paired or matched in some way.

As we begin this course, you should start with an awareness of the types of data we will be working with and learn to recognize situations which are more complex than those covered in this course.

Variables

> **Note**
>
> The columns in a dataset (representing **variables**) are often grouped and labeled by their role in our analysis.

For example, in many studies involving people, we often collect **demographic** variables such as gender, age, race, ethnicity, socioeconomic status, marital status, and many more.

> **Note**
>
> The **role** a variable plays in our analysis must also be considered.
>
> - In studies where we wish to predict one variable using one or more of the remaining variables, the variable we wish to predict is commonly called the **response** variable, the **outcome** variable, or the **dependent variable**.
> - Any variable we are using to predict or explain differences in the outcome is commonly called an **explanatory variable**, an **independent variable**, a **predictor** variable, or a **covariate**.

Various Uses of the Term INDEPENDENT in Statistics

Note: The word "**independent**" is used in statistics in numerous ways. Be careful to understand in what way the words "independent" or "independence" (as well as dependent or dependence) are used when you see them used in the materials.

- Here we have discussed **independent observations** (also called cases, individuals, or subjects).
- We have also used the term **independent variable** as another term for our explanatory variables.
- Later we will learn the formal probability definitions of **independent events** and **dependent events**.
- And when comparing groups we will define **independent samples** and **dependent samples**.

CHAPTER OVERVIEW

Unit 1: Exploratory Data Analysis

CO-1: Describe the roles biostatistics serves in the discipline of public health.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Exploratory Data Analysis Introduction (2 videos, 7:04 total)

The Big Picture

Learning Objectives

LO 1.3: Identify and differentiate between the components of the Big Picture of Statistics

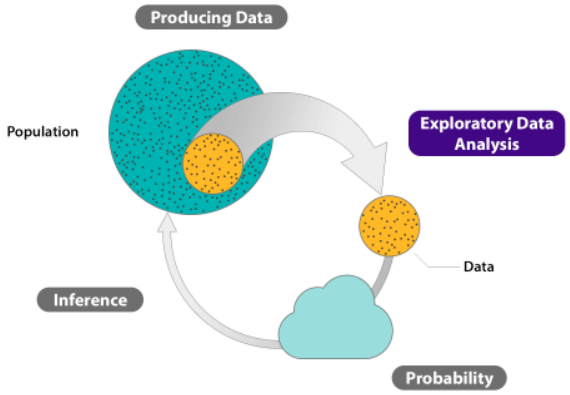
Recall "The Big Picture," the four-step process that encompasses statistics (as it is presented in this course):

1. Producing Data — Choosing a sample from the population of interest and collecting data.

2. Exploratory Data Analysis (EDA) {Descriptive Statistics} — Summarizing the data we've collected.

3. and 4. Probability and Inference — Drawing conclusions about the entire population based on the data collected from the sample.

Even though in practice it is the second step in the process, we are going to look at Exploratory Data Analysis (EDA) first. (If you have forgotten why, review the course structure information at the end of the page on The Big Picture and in the video covering The Big Picture.)



Exploratory Data Analysis

Learning Objectives

LO 1.5: Explain the uses and important features of exploratory data analysis.

As you can tell from the examples of datasets we have seen, raw data are not very informative. **Exploratory Data Analysis (EDA)** is how we make sense of the data by converting them from their raw form to a more informative one.

Note

In particular, **EDA consists of:**

- organizing and summarizing the raw data,
- discovering important features and patterns in the data and any striking deviations from those patterns, and then
- interpreting our findings in the context of the problem

And can be useful for:

- describing the distribution of a single variable (center, spread, shape, outliers)
- checking data (for errors or other problems)
- checking assumptions to more complex statistical analyses
- investigating relationships between variables

Exploratory data analysis (EDA) methods are often called **Descriptive Statistics** due to the fact that they simply describe, or provide estimates based on, the data at hand.

In Unit 4 we will cover methods of **Inferential Statistics** which use the results of a sample to make inferences about the population under study.

Comparisons can be visualized and values of interest estimated using EDA but descriptive statistics alone will provide no information about the certainty of our conclusions.

Important Features of Exploratory Data Analysis

There are two important features to the structure of the EDA unit in this course:

Note

- The material in this unit covers two broad topics:

  Examining Distributions — exploring data **one variable at a time**.

  Examining Relationships — exploring data **two variables at a time**.

Note

- In Exploratory Data Analysis, our exploration of data will always consist of the following two elements:

  visual displays, supplemented by

  numerical measures.

Try to remember these structural themes, as they will help you orient yourself along the path of this unit.

Examining Distributions

Learning Objectives

LO 6.1: Explain the meaning of the term distribution in statistics.

We will begin the EDA part of the course by exploring (or looking at) **one variable at a time**.

As we have seen, the data for each variable consist of a long list of values (whether numerical or not), and are not very informative in that form.

In order to convert these raw data into useful information, we need to summarize and then examine the **distribution** of the variable.

Note

By **distribution** of a variable, we mean:

- what values the variable takes, and
- how often the variable takes those values.

We will first learn how to summarize and examine the distribution of a single categorical variable, and then do the same for a single quantitative variable.

Case C-C

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.20: Classify a data analysis situation involving two variables according to the "role-type classification."

Learning Objectives

LO 4.21: For a data analysis situation involving two variables, determine the appropriate graphical display(s) and/or numerical measures(s) that should be used to summarize the data.

Video

Video: Case C-C (10:34)

Related SAS Tutorials

- 6A – (3:07) Two-Way (Contingency) Tables – EDA

Related SPSS Tutorials

- 6A – (7:57) Two-Way (Contingency) Tables – EDA

Two Categorical Variables

Recall the role-type classification table for framing our discussion about the relationship between two variables:

		Response	
		Categorical	Quantitative
Explanatory	Categorical	**C → C**	✓C → Q
	Quantitative	Q → C	Q → Q

We are done with case C → Q, and will now move on to case C → C, where we examine the relationship between two categorical variables.

Earlier in the course, (when we discussed the distribution of a **single** categorical variable) we examined the data obtained when a random sample of 1,200 U.S. college students were asked about their body image (underweight, overweight, or about right). We are now returning to this example, to address the following question:

If we had separated our sample of 1,200 U.S. college students by gender and looked at **males and females separately**, would we have found a similar distribution across body-image categories? More specifically, are men and women just as likely to think their weight is about right? Among those students who do not think their weight is about right, is there a difference between the genders in feelings about body image?

Answering these questions requires us to **examine the relationship between two categorical variables**, gender and body image. Because the question of interest is whether there is a gender effect on body image,

- the **explanatory** variable is **gender**, and
- the **response** variable is **body image**.

Here is what the raw data look like when we include the gender of each student:

Explanatory Response

Student	Gender	Body Image
.	.	.
.	.	.
student 25	M	overweight
student 26	M	about right
student 27	F	underweight
student 28	F	about right
student 29	M	about right
.	.	.
.	.	.

Once again the raw data is a long list of 1,200 genders and responses, and thus not very useful in that form.

Contingency Tables

Learning Objectives

LO 4.22: Define and explain the process of creating a contingency table (two-way table).

To start our exploration of how body image is related to gender, we need an informative display that summarizes the data. In order to summarize the relationship between two categorical variables, we create a display called a **two-way table** or **contingency table**.

Here is the two-way table for our example:

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

The table has the possible genders in the rows, and the possible responses regarding body image in the columns. At each intersection between row and column, we put the counts for how many times that combination of gender and body image occurred in the data. We sum across the rows to fill in the Total column, and we sum across the columns to fill in the Total row.

Complete the following activities related to this data.

Learn By Doing: Case C-C

Comments:

Note that from the way the two-way table is constructed, the Total row or column is a summary of one of the two categorical variables, ignoring the other. In our example:

- The Total row gives the summary of the categorical variable body image:

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

- The Total column gives the summary of the categorical variable gender:(These are the same counts we found earlier in the course when we looked at the single categorical variable body image, and did not consider gender.)

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

Finding Conditional (Row and Column) Percents

Learning Objectives

LO 4.23: Given a contingency table (two-way table), interpret the information it reveals about the association between two categorical variables by calculating and comparing conditional percentages.

So far we have organized the raw data in a much more informative display — the two-way table:

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

Remember, though, that our primary goal is to explore how body image is related to gender. Exploring the relationship between two categorical variables (in this case body image and gender) amounts to comparing the distributions of the response variable (in this case body image) across the different values of the explanatory variable (in this case males and females):

Compare these distributions! →

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

Note that it doesn't make sense to compare raw counts, because there are more females than males overall. So for example, it is not very informative to say "there are 560 females who responded 'about right' compared to only 295 males," since the 560 females are out of a total of 760, and the 295 males are out of a total of only 440.

We need to supplement our display, the two-way table, with some numerical measures that will allow us to compare the distributions. These numerical measures are found by simply **converting the counts to percents within (or restricted to) each value of the explanatory variable separately.**

In our example: We look at each gender separately, and convert the counts to percents **within that gender.** Let's start with females:

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560/760 = 73.7%	163/760 = 21.4%	37/760 = 4.9%	760/760 = 100%
	Male	%	%	%	%

Note that each count is converted to percents by dividing by the total number of females, 760. These numerical measures are called **conditional percents**, since we find them by "conditioning" on one of the genders.

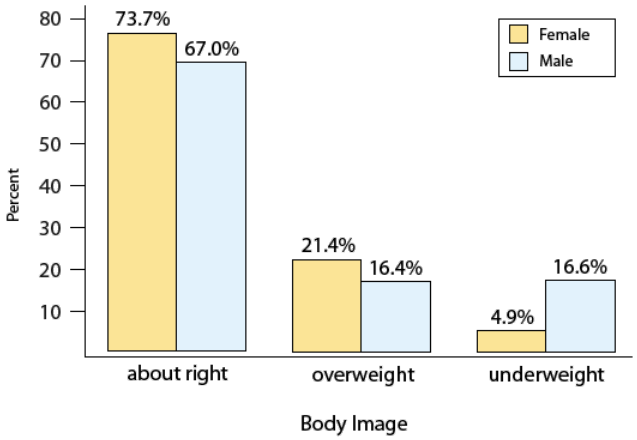
Now complete the following activities to calculate the row percentages for males.

Learn By Doing: Calculating Row Percentages

Comments:

- In our example, we chose to organize the data with the explanatory variable gender in rows and the response variable body image in columns, and thus our conditional percents were **row percents**, calculated within each row separately. Similarly, if the explanatory variable happens to sit in columns and the response variable in rows, our conditional percents will be **column percents**, calculated within each column separately. For an example, see the "Did I Get This?" exercises below.
- Another way to visualize the conditional percents, instead of a table, is the **double bar chart.** This display is quite common in newspapers.

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	73.7%	21.4%	4.9%	100%
	Male	67.0%	16.4%	16.6%	100%



Now that we have summarized the relationship between the categorical variables gender and body image, let's go back and interpret the results in the context of the questions that we posed.

Learn By Doing: Interpretation in Case C-C

Learn By Doing: Case C-C (Software)

For additional practice complete the following activities.

Did I Get This?: Case C-C

Let's Summarize

- The relationship between two categorical variables is summarized using:
  - **Data display:** two-way table, supplemented by
  - **Numerical measures:** conditional percentages.
- Conditional percentages are calculated for each value of the explanatory variable separately. They can be row percentages, if the explanatory variable "sits" in the rows, or column percentages, if the explanatory variable "sits" in the columns.
- When we try to understand the relationship between two categorical variables, we compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

Case C-Q

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.20: Classify a data analysis situation involving two variables according to the "role-type classification."

Learning Objectives

LO 4.21: For a data analysis situation involving two variables, determine the appropriate graphical display(s) and/or numerical measures(s) that should be used to summarize the data.

Video

Video: Case C-Q (6:34)

Related SAS Tutorials

- 7A (2:32) Numeric Summaries by Groups
- 7B (3:03) Side-By-Side Boxplots

Related SPSS Tutorials

- 7A (3:29) Numeric Summaries by Groups
- 7B (1:59) Side-By-Side Boxplots

Categorical Explanatory and Quantitative Response

Learning Objectives

LO 4.18: Compare and contrast distributions (of quantitative data) from two or more groups, and produce a brief summary, interpreting your findings in context.

Recall the role-type classification table for framing our discussion about the relationship between two variables:

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C → C	**C → Q**
	Quantitative	Q → C	Q → Q

We are now ready to start with Case C → Q, exploring the relationship between two variables where the explanatory variable is categorical, and the response variable is quantitative. As you'll discover, exploring relationships of this type is something we've already discussed in this course, but we didn't frame the discussion this way.

EXAMPLE: Hot Dogs

Background: People who are concerned about their health may prefer hot dogs that are low in calories. A study was conducted by a concerned health group in which 54 major hot dog brands were examined, and their calorie contents recorded. In addition, each brand was classified by type: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). The purpose of the study was to examine whether the **number of calories** a hot dog has is related to (or affected by) its **type**.

(Reference: Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics. Original source: Consumer Reports, June 1986, pp. 366-367.)

Answering this question requires us to examine the relationship between the categorical variable, Type and the quantitative variable Calories. Because the question of interest is whether the type of hot dog affects calorie content,

- the **explanatory** variable is **Type**, and
- the **response** variable is **Calories**.

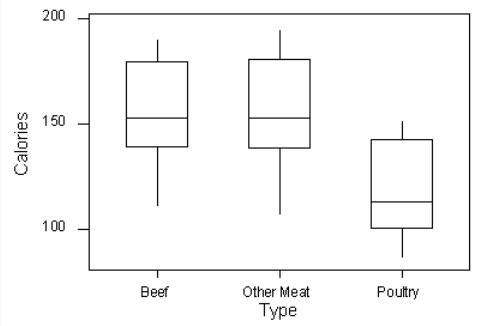
Here is what the raw data look like:

Explanatory ↖ Response ↗

	Type	Calories
Brand 1	Beef	186
Brand 2	Poultry	129
Brand 3	Beef	181
Brand 4	Meat	173
.	.	.
.	.	.
.	.	.
Brand 54	Poultry	144

The raw data are a list of types and calorie contents, and are not very useful in that form. To explore how the number of calories is related to the type of hot dog, we need an informative visual display of the data that will compare the three types of hot dogs with respect to their calorie content.

The visual display that we'll use is **side-by-side boxplots** (which we've seen before). The side-by-side boxplots will allow us to **compare the distribution** of calorie counts within each category of the explanatory variable, hot dog type:



As before, we supplement the side-by-side boxplots with the descriptive statistics of the calorie content (response) for each type of hot dog separately (i.e., for each level of the explanatory variable separately):

Let's summarize the results we obtained and interpret them in the context of the question we posed:

Statistic	Beef	Meat	Poultry
min	111	107	86
Q1	139.5	138.5	100.5
Median	152.5	153	113
Q3	179.75	180.5	142.5
Max	190	195	152

By examining the three side-by-side boxplots and the numerical measures, we see at once that poultry hot dogs, as a group, contain fewer calories than those made of beef or meat. The median number of calories in poultry hot dogs (113) is less than the median (and even the first quartile) of either of the other two distributions (medians 152.5 and 153). The spread of the three distributions is about the same, if IQR is considered (all slightly above 40), but the (full) ranges vary slightly more (beef: 80, meat: 88, poultry: 66). The general recommendation to the health-conscious consumer is to eat poultry hot dogs. It should be

noted, though, that since each of the three types of hot dogs shows quite a large spread among brands, simply buying a poultry hot dog does not guarantee a low-calorie food.

What we learn from this example is that when exploring the relationship between a categorical explanatory variable and a quantitative response (Case C→Q), we essentially **compare the distributions of the quantitative response for each category of the explanatory variable** using side-by-side boxplots supplemented by descriptive statistics. Recall that we have actually done this before when we talked about the boxplot and argued that boxplots are most useful when presented side by side for comparing distributions of two or more groups. This is exactly what we are doing here!

Here is another example:

EXAMPLE: SSHA

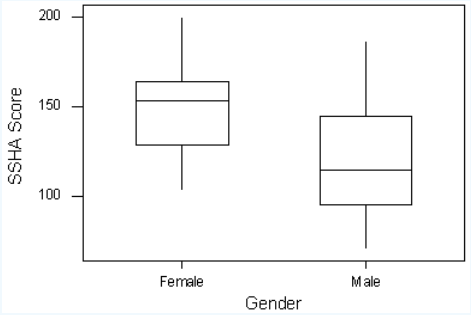
Background: The Survey of Study Habits and Attitudes (SSHA) is a psychological test designed to measure the motivation, study habits, and attitudes toward learning of college students. Is there a relationship between **gender** and **SSHA** scores? In other words, is there a "gender effect" on SSHA scores? Data were collected from 40 randomly selected college students, and here is what the raw data look like:

Explanatory ↖ Response ↗

	Gender	SSHA score
Student 1	Female	154
Student 2	Female	109
Student 3	Male	108
Student 4	Female	115
.	.	.
.	.	.
.	.	.
Student 40	Male	140

(Reference: Moore and McCabe. (2003). Introduction to the Practice of Statistics)

Side-by-side boxplots supplemented by descriptive statistics allow us to compare the distribution of SSHA scores within each category of the explanatory variable—gender:



Statistic	Female	Male
min	103	70
Q1	128.75	95
Median	153	114.5
Q3	163.75	144.5
Max	200	187

Let's summarize our results and interpret them:

By examining the side-by-side boxplots and the numerical measures, we see that in general females perform better on the SSHA than males. The median SSHA score of females is higher than the median score for males (153 vs. 114), and in fact, it is

even higher than the third quartile of the males' distribution (144.5). On the other hand, the males' scores display more variability, both in terms of IQR (49.5 vs. 35) and in terms of the full range of scores (117 vs. 97). Based on these results, it seems that there is a gender effect on SSHA score. It should be noted, though, that our sample consists of only 20 males and 20 females, so we should be cautious about making any kind of generalizations beyond this study. One interesting question that comes to mind is, "Why did we observe this relationship between gender and SSHA scores?" In other words, is there maybe an explanation for why females score higher on the SSHA? Let's leave it to the psychologists to try and answer that one.

Let's Summarize

- The relationship between a categorical explanatory variable and a quantitative response variable is summarized using:
  - **Visual display:** side-by-side boxplots
  - **Numerical measures:** descriptive statistics used for one quantitative variable calculated in each group
- Exploring the relationship between a categorical explanatory variable and a quantitative response variable amounts to comparing the distributions of the quantitative response for each category of the explanatory variable. In particular, we look at how the distribution of the response variable differs between the values of the explanatory variable

Case Q-Q

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.20: Classify a data analysis situation involving two variables according to the "role-type classification."

Learning Objectives

LO 4.21: For a data analysis situation involving two variables, determine the appropriate graphical display(s) and/or numerical measures(s) that should be used to summarize the data.

Video

Video: Case Q-Q (2:30)

Related SAS Tutorials

- 9A – (3:53) Basic Scatterplots
- 9B – (2:29) Grouped Scatterplots
- 9C – (3:46) Pearson's Correlation Coefficient
- 9D – (3:00) Simple Linear Regression – EDA

Related SPSS Tutorials

- 9A – (2:38) Basic Scatterplots
- 9B – (2:54) Grouped Scatterplots
- 9C – (3:35) Pearson's Correlation Coefficient
- 9D – (2:53) Simple Linear Regression – EDA

Introduction – Two Quantitative Variables

Here again is the role-type classification table for framing our discussion about the relationship between two variables:

		Response	
		Categorical	Quantitative
Explanatory	Categorical	✓C→C	✓C→Q
	Quantitative	Q→C	**Q→Q**

Before reading further, try this interactive online data analysis applet.

Interactive Applet: Case Q-Q

We are done with cases C→Q and C→C, and now we will move on to case Q→Q, where we examine the relationship between two quantitative variables.

In this section we will discuss scatterplots, which are the appropriate visual display in this case along with numerical methods for linear relationships including correlation and linear regression.

Scatterplots

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.21: For a data analysis situation involving two variables, determine the appropriate graphical display(s) and/or numerical measures(s) that should be used to summarize the data.

Video

Video: Scatterplots (7:20)

Related SAS Tutorials

- 9A – (3:53) Basic Scatterplots
- 9B – (2:29) Grouped Scatterplots
- 9C – (3:46) Pearson's Correlation Coefficient
- 9D – (3:00) Simple Linear Regression – EDA

Related SPSS Tutorials

- 9A – (2:38) Basic Scatterplots
- 9B – (2:54) Grouped Scatterplots
- 9C – (3:35) Pearson's Correlation Coefficient
- 9D – (2:53) Simple Linear Regression – EDA

In the previous two cases we had a categorical explanatory variable, and therefore exploring the relationship between the two variables was done by comparing the distribution of the response variable for each category of the explanatory variable:

- In case C→Q we compared distributions of the quantitative response.
- In case C→C we compared distributions of the categorical response.

Case Q→Q is different in the sense that both variables (in particular the explanatory variable) are quantitative. As you will discover, although we are still in essence comparing the distribution of one variable for different values of the other, this case will require a different kind of treatment and tools.

Learning Objectives

LO 4.24: Explain the process of creating a scatterplot.

Creating Scatterplots

Let's start with an example:

EXAMPLE: Highway Signs

A Pennsylvania research firm conducted a study in which 30 drivers (of ages 18 to 82 years old) were sampled, and for each one, the maximum distance (in feet) at which he/she could read a newly designed sign was determined. The goal of this study was to explore the relationship between a driver's **age** and the **maximum distance** at which signs were legible, and then use the study's findings to improve safety for older drivers. (Reference: Utts and Heckard, *Mind on Statistics* (2002). Original source: Data collected by Last Resource, Inc, Bellfonte, PA.)

Since the purpose of this study is to explore the effect of age on maximum legibility distance,

- the **explanatory** variable is **Age**, and

- the **response** variable is **Distance**.

Here is what the raw data look like:

Explanatory ↖ Response ↗

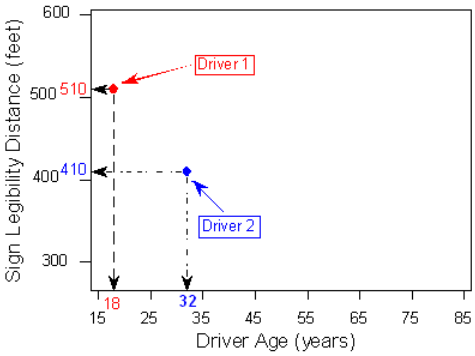
	Age	Distance
Driver 1	**18**	**510**
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

Note that the data structure is such that for each individual (in this case driver 1….driver 30) we have a pair of values (in this case representing the driver's age and distance). We can therefore think about these data as 30 pairs of values: (18, 510), (32, 410), (55, 420), … , (82, 360).

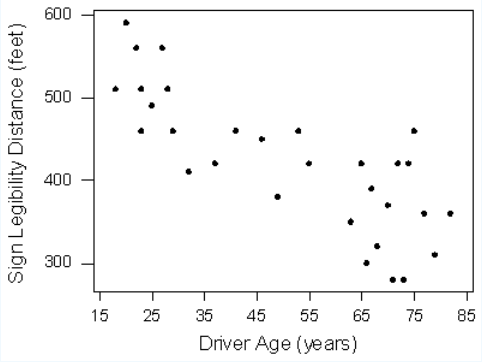
The first step in exploring the relationship between driver age and sign legibility distance is to create an appropriate and informative graphical display. The appropriate graphical display for examining the relationship between two quantitative variables is the **scatterplot**. Here is how a scatterplot is constructed for our example:

To create a scatterplot, each pair of values is plotted, so that the value of the explanatory variable (X) is plotted on the horizontal axis, and the value of the response variable (Y) is plotted on the vertical axis. In other words, each individual (driver, in our example) appears on the scatterplot as a single point whose X-coordinate is the value of the explanatory variable for that individual, and whose Y-coordinate is the value of the response variable. Here is an illustration:

	Age (X)	Distance (Y)
Driver 1	**18**	**510**
Driver 2	**32**	**410**
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360



And here is the completed scatterplot:



Comment:

- It is important to mention again that when creating a scatterplot, the explanatory variable should always be plotted on the horizontal X-axis, and the response variable should be plotted on the vertical Y-axis. If in a specific example we do not have a clear distinction between explanatory and response variables, each of the variables can be plotted on either axis.

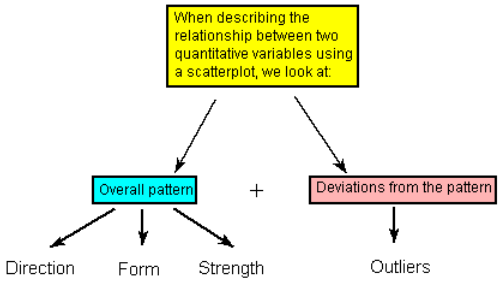
Interpreting Scatterplots

Learning Objectives

LO 4.25: Describe the relationship displayed in a scatterplot including: a) the overall pattern, b) striking deviations from the pattern.

How do we explore the relationship between two quantitative variables using the scatterplot? What should we look at, or pay attention to?

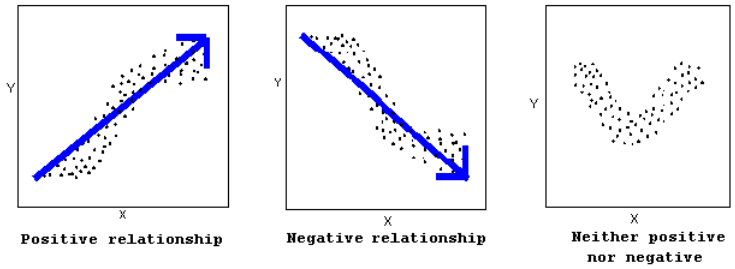
Recall that when we described the distribution of a single quantitative variable with a histogram, we described the overall pattern of the distribution (shape, center, spread) and any deviations from that pattern (outliers). **We do the same thing with the scatterplot.** The following figure summarizes this point:



As the figure explains, when describing the **overall pattern** of the relationship we look at its direction, form and strength.

Direction

- The **direction** of the relationship can be positive, negative, or neither:



A **positive (or increasing) relationship** means that an increase in one of the variables is associated with an increase in the other.

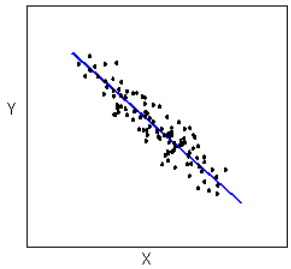
A **negative (or decreasing) relationship** means that an increase in one of the variables is associated with a decrease in the other.

Not all relationships can be classified as either positive or negative.

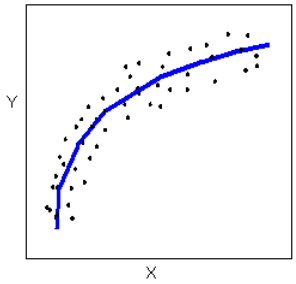
Form

- The **form** of the relationship is its general shape. When identifying the form, we try to find the simplest way to describe the shape of the scatterplot. There are many possible forms. Here are a couple that are quite common:

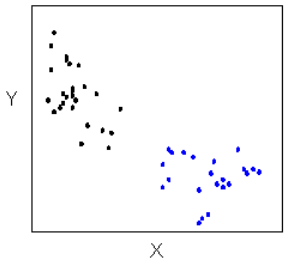
Relationships with a **linear** form are most simply described as points scattered about a line:



Relationships with a **non-linear (sometimes called curvilinear)** form are most simply described as points dispersed around the same curved line:

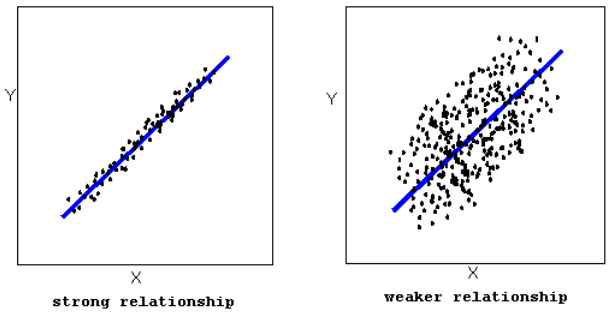


There are many other possible forms for the relationship between two quantitative variables, but linear and curvilinear forms are quite common and easy to identify. Another form-related pattern that we should be aware of is clusters in the data:



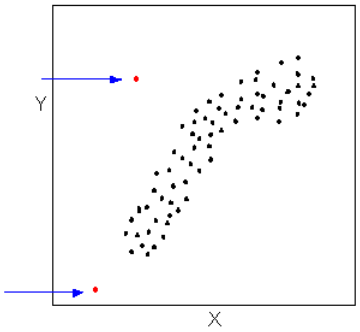
Strength

- The **strength** of the relationship is determined by how closely the data follow the form of the relationship. Let's look, for example, at the following two scatterplots displaying positive, linear relationships:



The strength of the relationship is determined by how closely the data points follow the form. We can see that in the left scatterplot the data points follow the linear pattern quite closely. This is an example of a strong relationship. In the right scatterplot, the points also follow the linear pattern, but much less closely, and therefore we can say that the relationship is weaker. In general, though, assessing the strength of a relationship just by looking at the scatterplot is quite problematic, and we need a numerical measure to help us with that. We will discuss that later in this section.

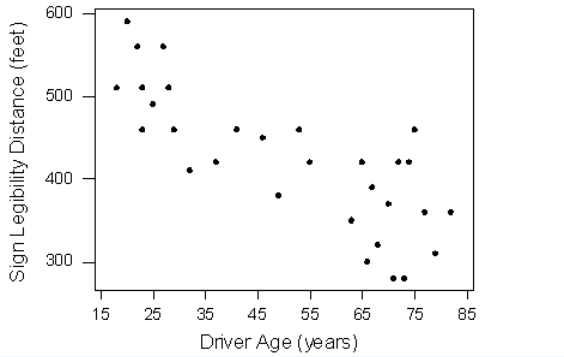
- Data points that **deviate from the pattern** of the relationship are called **outliers**. We will see several examples of outliers during this section. Two outliers are illustrated in the scatterplot below:



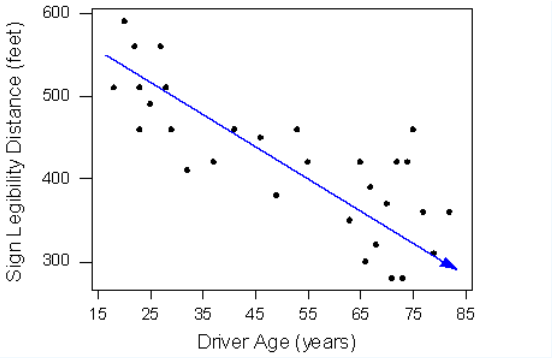
Let's go back now to our example, and use the scatterplot to examine the relationship between the age of the driver and the maximum sign legibility distance.

EXAMPLE: Highway Signs

Here is the scatterplot:



The direction of the relationship is **negative**, which makes sense in context, since as you get older your eyesight weakens, and in particular older drivers tend to be able to read signs only at lesser distances. An arrow drawn over the scatterplot illustrates the negative direction of this relationship:



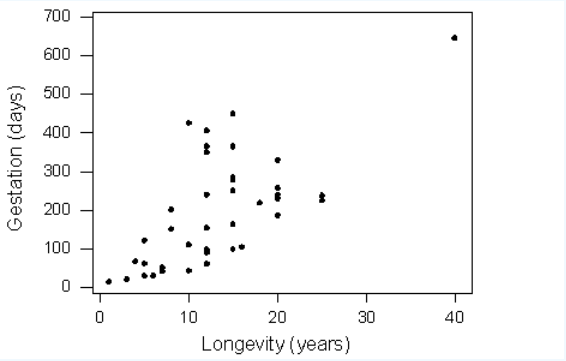
The form of the relationship seems to be **linear**. Notice how the points tend to be scattered about the line. Although, as we mentioned earlier, it is problematic to assess the strength without a numerical measure, the relationship appears to be **moderately strong**, as the data is fairly tightly scattered about the line. Finally, all the data points seem to "obey" the pattern — there **do not appear to be any outliers**.

We will now look at two more examples:

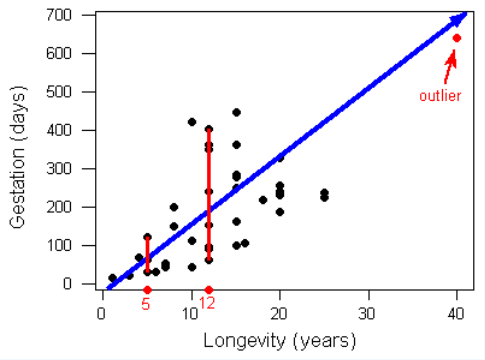
EXAMPLE: Average Gestation Period

The average gestation period, or time of pregnancy, of an animal is closely related to its longevity (the length of its lifespan). Data on the average gestation period and longevity (in captivity) of 40 different species of animals have been examined, with the purpose of examining how the gestation period of an animal is related to (or can be predicted from) its longevity. (Source: Rossman and Chance. (2001). Workshop statistics: Discovery with data and Minitab. Original source: The 1993 world almanac and book of facts).

Here is the scatterplot of the data.



What can we learn about the relationship from the scatterplot? The direction of the relationship is **positive**, which means that animals with longer life spans tend to have longer times of pregnancy (this makes intuitive sense). An arrow drawn over the scatterplot below illustrates this:



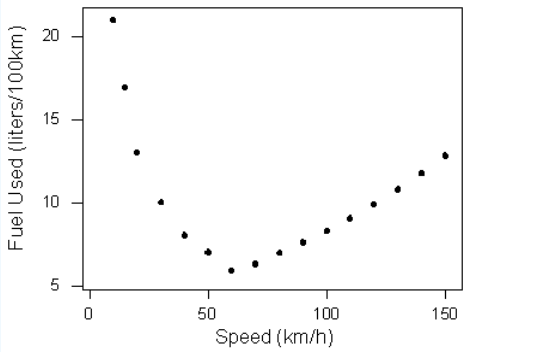
The form of the relationship is again essentially **linear**. There appears to be **one outlier**, indicating an animal with an exceptionally long longevity and gestation period. (This animal happens to be the elephant.) Note that while this outlier definitely deviates from the rest of the data in term of its magnitude, it **does** follow the direction of the data.

Comment:

- Another feature of the scatterplot that is worth observing is how the variation in gestation increases as longevity increases. This fact is illustrated by the two red vertical lines at the bottom left part of the graph. Note that the gestation periods for animals that live 5 years range from about 30 days up to about 120 days. On the other hand, the gestation periods of animals that live 12 years vary much more, and range from about 60 days up to more than 400 days.

EXAMPLE: Fuel Usage

As a third example, consider the relationship between the average amount of fuel used (in liters) to drive a fixed distance in a car (100 kilometers), and the speed at which the car is driven (in kilometers per hour). (Source: Moore and McCabe, (2003). Introduction to the practice of statistics. Original source: T.N. Lam. (1985). "Estimating fuel consumption for engine size," Journal of Transportation Engineering, vol. 111)

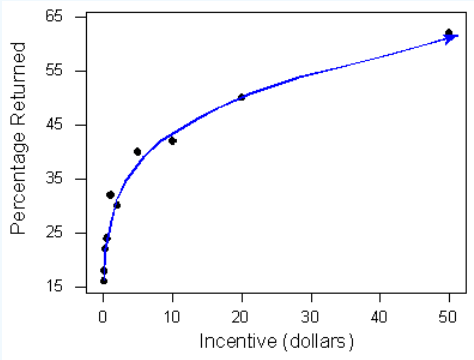


The data describe a relationship that decreases and then increases — the amount of fuel consumed decreases rapidly to a minimum for a car driving 60 kilometers per hour, and then increases gradually for speeds exceeding 60 kilometers per hour. This suggests that the speed at which a car economizes on fuel the most is about 60 km/h. This forms a non-linear (curvilinear) relationship that seems to be very strong, as the observations seem to perfectly fit the curve. Finally, there do not appear to be any outliers.

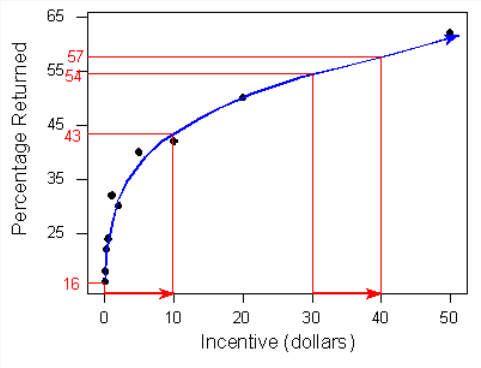
Learn By Doing: Scatterplots

EXAMPLE: Return on Incentives

The example in the last activity provides a great opportunity for interpretation of the form of the relationship in context. Recall that the example examined how the percentage of participants who completed a survey is affected by the monetary incentive that researchers promised to participants. Here again is the scatterplot that displays the relationship:



The positive relationship definitely makes sense in context, but what is the interpretation of the non-linear (curvilinear) form in the context of the problem? How can we explain (in context) the fact that the relationship seems at first to be increasing very rapidly, but then slows down? The following graph will help us:



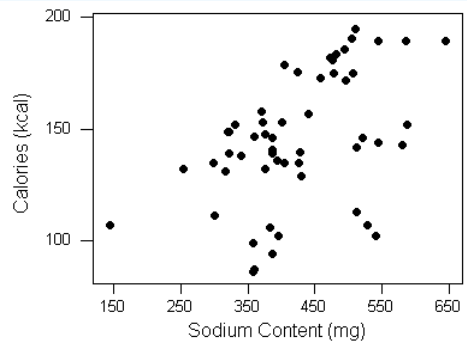
Note that when the monetary incentive increases from \$0 to \$10, the percentage of returned surveys increases sharply — an increase of 27% (from 16% to 43%). However, the same increase of \$10 from \$30 to \$40 doesn't result in the same dramatic increase in the percentage of returned surveys — it results in an increase of only 3% (from 54% to 57%). The form displays the phenomenon of "diminishing returns" — a return rate that after a certain point fails to increase proportionately to additional outlays of investment. \$10 is worth more to people relative to \$0 than \$30 is relative to \$10.

A Labeled (or Grouped) Scatterplot

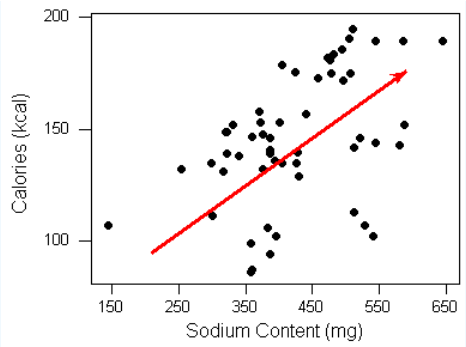
In certain circumstances, it may be reasonable to indicate different subgroups or categories within the data on the scatterplot, by labeling each subgroup differently. The result is sometimes called a **labeled scatterplot** or **grouped scatterplot**, and can provide further insight about the relationship we are exploring. Here is an example.

✔ EXAMPLE: Hot Dogs

The scatterplot below displays the relationship between the sodium and calorie content of 54 brands of hot dogs. Note that in this example there is no clear explanatory-response distinction, and we decided to have sodium content as the explanatory variable, and calorie content as the response variable.

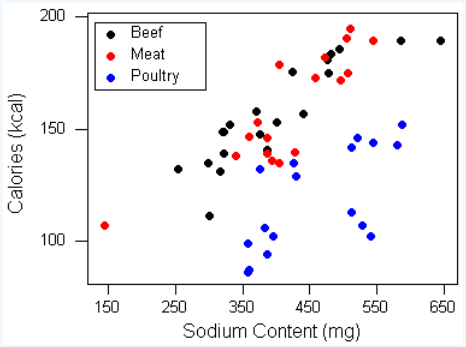


The scatterplot displays a positive relationship, which means that hot dogs containing more sodium tend to be higher in calories.

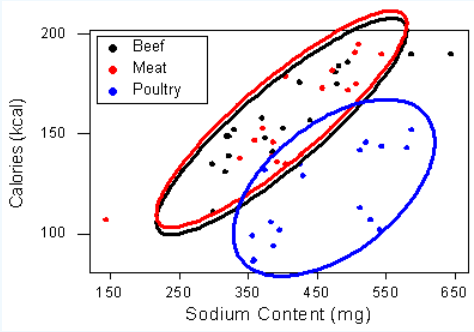


The form of the relationship, however, is kind of hard to determine. Maybe if we label the scatterplot, indicating the type of hot dogs, we will get a better understanding of the form.

Here is the labeled scatterplot, with the three different colors representing the three types of hot dogs, as indicated.

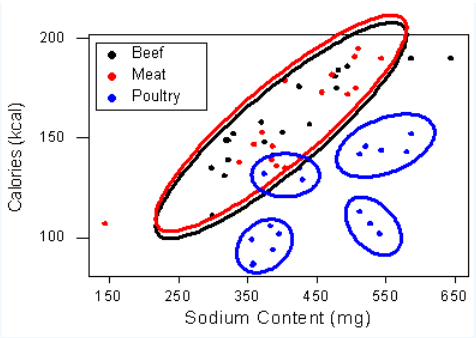


The display does give us more insight about the form of the relationship between sodium and calorie content.



It appears that there is a positive relationship within all three types. In other words, we can generally expect hot dogs that are higher in sodium to be higher in calories, no matter what type of hot dog we consider. In addition, we can see that hot dogs made of poultry (indicated in blue) are generally lower in calories. This is a result we have seen before.

Interestingly, it appears that the form of the relationship specifically for poultry is further clustered, and we can only speculate about whether there is another categorical variable that describes these apparent sub-categories of poultry hot dogs.



Learn By Doing: Scatterplots (Software)

Let's Summarize

- The relationship between two quantitative variables is visually displayed using the **scatterplot**, where each point represents an individual. We always plot the explanatory variable on the horizontal X axis, and the response variable on the vertical Y axis.
- When we explore a relationship using the scatterplot we should describe the **overall pattern** of the relationship and any **deviations** from that pattern. To describe the overall pattern consider the **direction**, **form** and **strength** of the relationship. Assessing the strength just by looking at the scatterplot can be problematic; using a numerical measure to determine strength will be discussed later in this course.
- Adding labels to the scatterplot that indicate different groups or categories within the data might help us get more insight about the relationship we are exploring.

Linear Relationships – Correlation

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.21: For a data analysis situation involving two variables, determine the appropriate graphical display(s) and/or numerical measures(s) that should be used to summarize the data.

Video

Video: Linear Relationships – Correlation (8:37)

Related SAS Tutorials

- 9A – (3:53) Basic Scatterplots
- 9B – (2:29) Grouped Scatterplots
- 9C – (3:46) Pearson's Correlation Coefficient
- 9D – (3:00) Simple Linear Regression – EDA

Related SPSS Tutorials

- 9A – (2:38) Basic Scatterplots
- 9B – (2:54) Grouped Scatterplots
- 9C – (3:35) Pearson's Correlation Coefficient
- 9D – (2:53) Simple Linear Regression – EDA

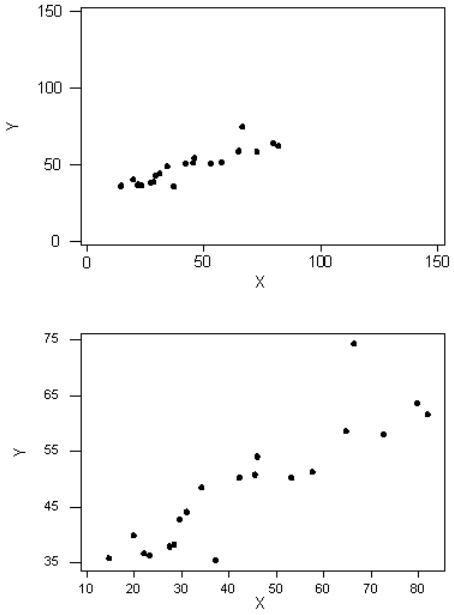
Introduction

So far we have visualized relationships between two quantitative variables using scatterplots, and described the overall pattern of a relationship by considering its direction, form, and strength. We noted that assessing the strength of a relationship just by looking at the scatterplot is quite difficult, and therefore we need to supplement the scatterplot with some kind of numerical measure that will help us assess the strength.

In this part, we will restrict our attention to the **special case of relationships that have a linear form**, since they are quite common and relatively simple to detect. More importantly, there exists a numerical measure that assesses the strength of the **linear** relationship between two quantitative variables with which we can supplement the scatterplot. We will introduce this numerical measure here and discuss it in detail.

Even though from this point on we are going to focus only on **linear** relationships, it is important to remember that **not every relationship between two quantitative variables has a linear form.** We have actually seen several examples of relationships that are not linear. The statistical tools that will be introduced here are **appropriate only for examining linear relationships,** and as we will see, when they are used in nonlinear situations, these tools can lead to errors in reasoning.

Let's start with a motivating example. Consider the following two scatterplots.



We can see that in both cases, the direction of the relationship is **positive** and the form of the relationship is **linear**. What about the strength? Recall that the strength of a relationship is the extent to which the data follow its form.

Learn By Doing: Strength of Correlation

The purpose of this example was to illustrate how assessing the strength of the **linear** relationship from a scatterplot alone is problematic, since our judgment might be affected by the scale on which the values are plotted. This example, therefore, provides a motivation for the **need** to supplement the scatterplot with a **numerical measure** that will **measure the strength** of the **linear** relationship between two quantitative variables.

The Correlation Coefficient — *r*

Learning Objectives

LO 4.26: Explain the limitations of Pearson's correlation coefficient (r) as a measure of the association between two quantitative variables.

Learning Objectives

LO 4.27: In the special case of a linear relationship, interpret Pearson's correlation coefficient (*r*) in context.

The numerical measure that assesses the strength of a **linear** relationship is called the **correlation coefficient**, and is denoted by r. We will:

- give a definition of the correlation r,
- discuss the calculation of r,
- explain how to interpret the value of r, and
- talk about some of the properties of r.

Correlation Coefficient: The correlation coefficient (r) is a numerical measure that measures the **strength** and **direction** of a **linear** relationship between two quantitative variables.

Calculation: r is calculated using the following formula:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

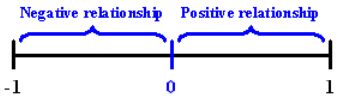
However, the calculation of the correlation (r) is not the focus of this course. We will use a statistics package to calculate r for us, and the **emphasis** of this course will be on the **interpretation** of its value.

Interpretation

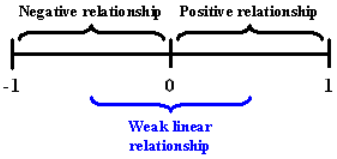
Once we obtain the value of r, its interpretation with respect to the strength of **linear** relationships is quite simple, as these images illustrate:

The value of r ranges from -1 to 1.

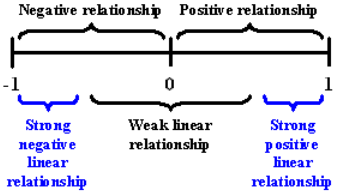
Negative values of r indicate a negative direction for a linear relationship, and positive values of r indicate a positive direction for a linear relationship.



Values of r that are close to 0—either negative or positive—indicate a weak linear relationship.



Values that are close to -1 or close to 1 indicate a strong linear relationship, either negative or positive.



In order to get a better sense for how the value of *r* relates to the strength of the **linear** relationship, take a look the following applets.

> **Interactive Applets:** Correlation

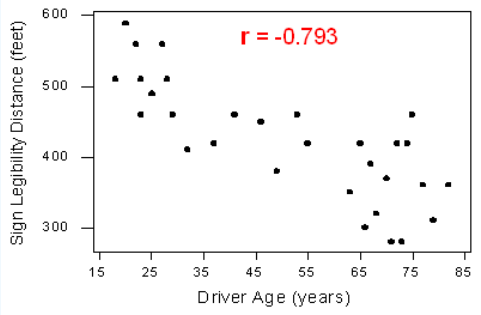
If you will be using correlation often in your research, I highly urge you to read the following more detailed discussion of correlation.

> **(Optional) Outside Reading:** Correlation Coefficients (≈ 2700 words)

Now that we understand the use of *r* as a numerical measure for assessing the direction and strength of **linear** relationships between quantitative variables, we will look at a few examples.

EXAMPLE: Highway Sign Visibility

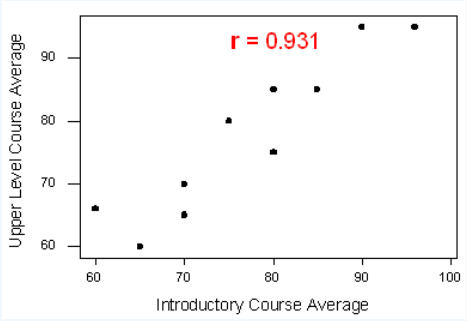
Earlier, we used the scatterplot below to find a **negative linear** relationship between the age of a driver and the maximum distance at which a highway sign was legible. What about the strength of the relationship? It turns out that the correlation between the two variables is r = -0.793.



Since r < 0, it confirms that the direction of the relationship is negative (although we really didn't need r to tell us that). Since r is relatively close to -1, it suggests that the relationship is moderately strong. In context, the negative correlation confirms that the maximum distance at which a sign is legible generally decreases with age. Since the value of r indicates that the **linear** relationship is moderately strong, but not perfect, we can expect the maximum distance to vary somewhat, even among drivers of the same age.

EXAMPLE: Statistic Courses

A statistics department is interested in tracking the progress of its students from entry until graduation. As part of the study, the department tabulates the performance of 10 students in an introductory course and in an upper-level course required for graduation. What is the relationship between the students' course averages in the two courses? Here is the scatterplot for the data:



The scatterplot suggests a relationship that is **positive** in direction, **linear** in form, and seems quite strong. The value of the correlation that we find between the two variables is r = 0.931, which is very close to 1, and thus confirms that indeed the **linear** relationship is very strong.

Comments:

- Note that in both examples we supplemented the scatterplot with the correlation (r). Now that we have the correlation (r), why do we still need to look at a scatterplot when examining the relationship between two quantitative variables?
- The **correlation** coefficient can **only** be interpreted as the **measure of the strength of a linear relationship**, so we need the scatterplot to verify that the relationship indeed looks **linear**. This point and its importance will be clearer after we examine a few properties of r.

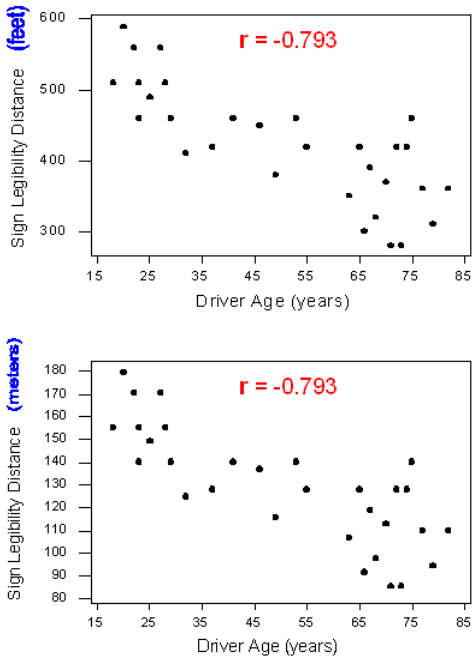
Did I Get This? Correlation Coefficient

Properties of *r*

We will now discuss and illustrate several important properties of the correlation coefficient as a numerical measure of the strength of a **linear** relationship.

- The correlation does not change when the units of measurement of either one of the variables change. In other words, if we **change the units of measurement** of the explanatory variable and/or the response variable, this has **no effect on the correlation (r)**.

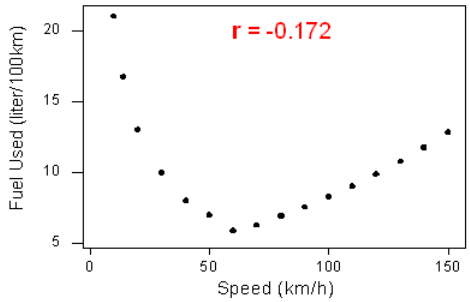
To illustrate this, below are two versions of the scatterplot of the relationship between sign legibility distance and driver's age:



The top scatterplot displays the original data where the maximum distances are measured **in feet**. The bottom scatterplot displays the same relationship, but with maximum distances changed to **meters**. Notice that the Y-values have changed, but the correlations are the same. This is an example of how changing the units of measurement of the response variable has no effect on r, but as we indicated above, the same is true for changing the units of the explanatory variable, or of both variables.

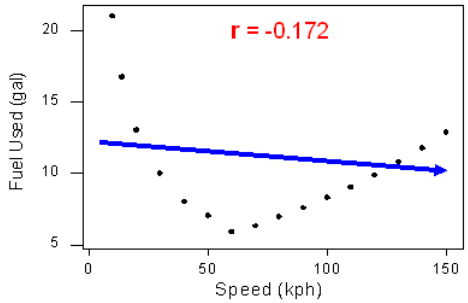
This might be a good place to comment that the correlation (r) is **"unitless"**. It is just a number.

- The correlation **only measures the strength of a linear relationship** between two variables. **It ignores any other type of relationship, no matter how strong it is.** For example, consider the relationship between the average fuel usage of driving a fixed distance in a car, and the speed at which the car drives:

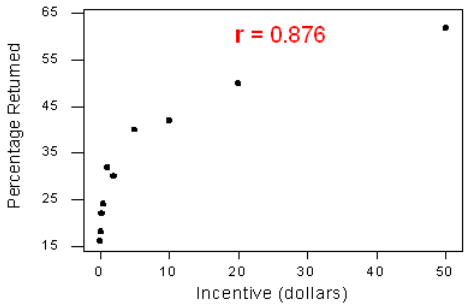


Our data describe a fairly simple non-linear (sometimes called curvilinear) relationship: the amount of fuel consumed decreases rapidly to a minimum for a car driving 60 kilometers per hour, and then increases gradually for speeds exceeding 60 kilometers per hour. The relationship is very strong, as the observations seem to perfectly fit the curve.

Although the relationship is strong, the correlation r = -0.172 indicates a weak **linear** relationship. This makes sense considering that the data fails to adhere closely to a linear form:



- The correlation by itself is **not** enough to determine whether or not a relationship is linear. To see this, let's consider the study that examined the effect of monetary incentives on the return rate of questionnaires. Below is the scatterplot relating the percentage of participants who completed a survey to the monetary incentive that researchers promised to participants, in which we find a **strong non-linear (sometimes called curvilinear) relationship:**



The relationship is non-linear (sometimes called curvilinear), yet the correlation r = 0.876 is quite close to 1.

In the last two examples we have seen two very strong non-linear (sometimes called curvilinear) relationships, one with a correlation close to 0, and one with a correlation close to 1. Therefore, the correlation alone does not indicate whether a relationship is **linear** or not. The important principle here is:

Always look at the data!

- The correlation is heavily influenced by outliers. As you will learn in the next two activities, the way in which the outlier influences the correlation depends upon whether or not the outlier is consistent with the pattern of the **linear** relationship.

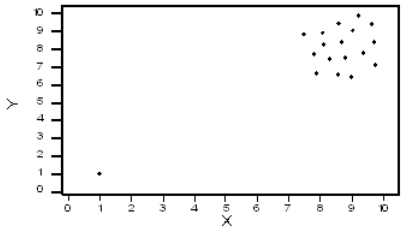
> **Interactive Applet:** Correlation and Outliers

Hopefully, you've noticed the correlation decreasing when you created this kind of outlier, which **is not consistent** with the pattern of the relationship.

The next activity will show you how an outlier that **is consistent** with the direction of the linear relationship actually strengthens it.

> **Learn By Doing:** Correlation and Outliers (Software)

In the previous activity, we saw an example where there was a positive **linear** relationship between the two variables, and including the outlier just "strengthened" it. Consider the hypothetical data displayed by the following scatterplot:



In this case, the low outlier gives an "illusion" of a positive **linear** relationship, whereas in reality, there is no **linear** relationship between X and Y.

Linear Relationships – Linear Regression

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.21: For a data analysis situation involving two variables, determine the appropriate graphical display(s) and/or numerical measures(s) that should be used to summarize the data.

Video

Video: Linear Relationships – Linear Regression (5:58)

Related SAS Tutorials

- 9A – (3:53) Basic Scatterplots
- 9B – (2:29) Grouped Scatterplots
- 9C – (3:46) Pearson's Correlation Coefficient
- 9D – (3:00) Simple Linear Regression – EDA

Related SPSS Tutorials

- 9A – (2:38) Basic Scatterplots
- 9B – (2:54) Grouped Scatterplots
- 9C – (3:35) Pearson's Correlation Coefficient
- 9D – (2:53) Simple Linear Regression – EDA

Summarizing the Pattern of the Data with a Line

Learning Objectives

LO 4.28: In the special case of a linear relationship, interpret the slope of the regression line and use the regression line to make predictions.

So far we've used the scatterplot to describe the relationship between two quantitative variables, and in the special case of a linear relationship, we have supplemented the scatterplot with the correlation (r).

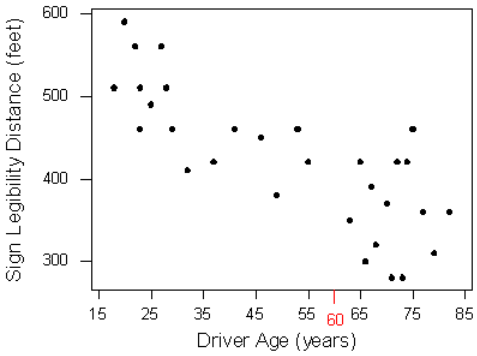
The correlation, however, doesn't fully characterize the linear relationship between two quantitative variables — it only measures the strength and direction. We often want to describe more precisely how one variable changes with the other (by "more precisely," we mean more than just the direction), or predict the value of the response variable for a given value of the explanatory variable.

In order to be able to do that, we need to summarize the linear relationship with a line that best fits the linear pattern of the data. In the remainder of this section, we will introduce a way to find such a line, learn how to interpret it, and use it (cautiously) to make predictions.

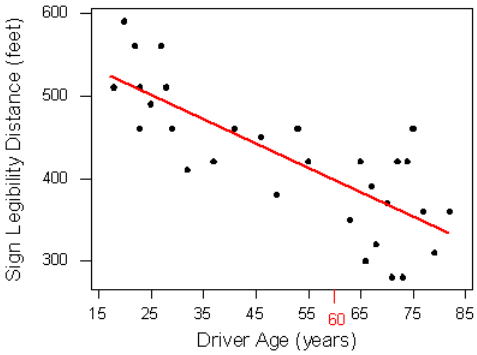
Again, let's start with a motivating example:

Earlier, we examined the linear relationship between the age of a driver and the maximum distance at which a highway sign was legible, using both a scatterplot and the correlation coefficient. Suppose a government agency wanted to predict the maximum distance at which the sign would be legible for 60-year-old drivers, and thus make sure that the sign could be used safely and effectively.

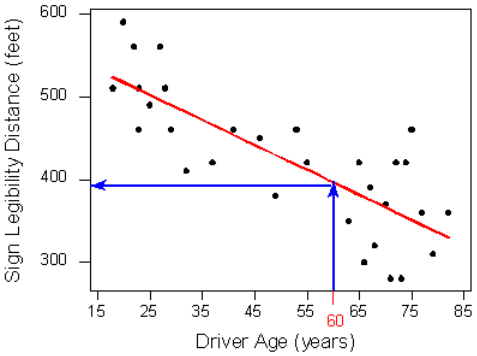
How would we make this prediction?



It would be useful if we could find a line (such as the one that is presented on the scatterplot) that represents the general pattern of the data, because then,



and predict that 60-year-old drivers could see the sign from a distance of just under 400 feet we would simply use this line to find the distance that corresponds to an age of 60 like this:

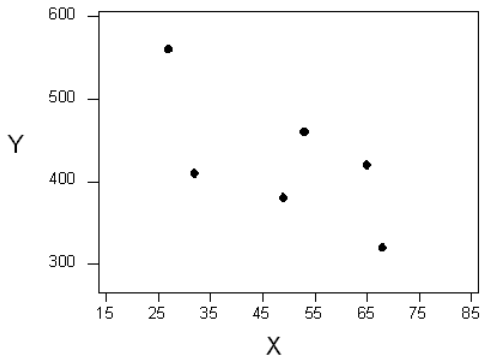


How and why did we pick this particular line (the one shown in red in the above walkthrough) to describe the dependence of the maximum distance at which a sign is legible upon the age of a driver? What line exactly did we choose? We will return to this example once we can answer that question with a bit more precision.

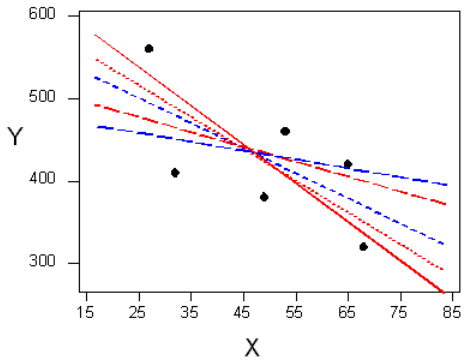
> **Interactive Applets:** Regression by Eye

The technique that specifies the dependence of the response variable on the explanatory variable is called **regression**. When that dependence is linear (which is the case in our examples in this section), the technique is called **linear regression**. Linear regression is therefore the technique of finding the line that best fits the pattern of the linear relationship (or in other words, the line that best describes how the response variable linearly depends on the explanatory variable).

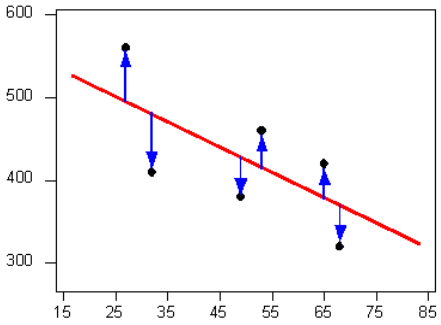
To understand how such a line is chosen, consider the following very simplified version of the age-distance example (we left just 6 of the drivers on the scatterplot):



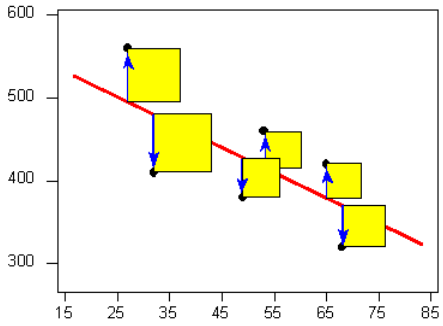
There are many lines that look like they would be good candidates to be the line that best fits the data:



It is doubtful that everyone would select the same line in the plot above. We need to agree on what we mean by "best fits the data"; in other words, we need to agree on a criterion by which we would select this line. We want the line we choose to be close to the data points. In other words, whatever criterion we choose, it had better somehow take into account the vertical deviations of the data points from the line, which are marked with blue arrows in the plot below:



The most commonly used criterion is called the **least squares** criterion. This criterion says: Among all the lines that look good on your data, choose the one that has the smallest sum of squared vertical deviations. Visually, each squared deviation is represented by the area of one of the squares in the plot below. Therefore, we are looking for the line that will have the smallest total yellow area.



This line is called the **least-squares regression line**, and, as we'll see, it fits the linear pattern of the data very well.

For the remainder of this lesson, you'll need to feel comfortable with the algebra of a straight line. In particular you'll need to be familiar with the **slope** and the **intercept** in the equation of a line, and their interpretation.

Many Students Wonder: Algebra Review – Linear Equation

Interactive Applet: Linear Equations – Effect of Changing the Slope or Intercept on the Line

Like any other line, the equation of the least-squares regression line for summarizing the linear relationship between the response variable (**Y**) and the explanatory variable (**X**) has the form: **Y = a + bX**

All we need to do is calculate the intercept ***a***, and the slope ***b***, which we will learn to do using software.

The **slope** of the least squares regression line can be interpreted as the estimated (or predicted) **change in the mean (or average) value of the response variable when the explanatory variable increases by 1 unit.**

EXAMPLE: Age-Distance

Let's revisit our age-distance example, and find the **least-squares regression line**. The following output will be helpful in getting the 5 values we need:

Column	n	Mean	Std. Dev.	Std. Err.	Min	Q1	Median	Q3	Max
Age	30	51	21.776293	3.9757888	18	28	54	71	82
Distance	30	423	82.802216	15.117547	280	360	420	460	590

- Dependent Variable: Distance
- Independent Variable: Age
- Correlation Coefficient (**r**) = -0.7929
- The **least squares regression line** for this example is:

$$\text{Distance} = 576 + (-3 * \text{Age}$$

- This means that for every 1-unit increase of the explanatory variable, there is, on average, a 3-unit decrease in the response variable. The interpretation **in context** of the slope (-3) is, therefore: In this dataset, when age increases by 1 year the **average** maximum distance at which subjects can read a sign is expected to **decrease by 3 feet.**
- Here is the regression line plotted on the scatterplot:

As we can see, the regression line fits the linear pattern of the data quite well.

Let's go back now to our motivating example, in which we wanted to predict the maximum distance at which a sign is legible for a 60-year-old. Now that we have found the least squares regression line, this prediction becomes quite easy:

EXAMPLE: Age-Distance

Practically, what the figure tells us is that in order to find the predicted legibility distance for a 60-year-old, we plug Age = 60 into the regression line equation, to find that:

Predicted distance = 576 + (- 3 * 60) = 396

396 feet is our best prediction for the maximum distance at which a sign is legible for a 60-year-old.

Did I Get This?: Linear Regression

Comment About Predictions:

- Suppose a government agency wanted to design a sign appropriate for an even wider range of drivers than were present in the original study. They want to predict the maximum distance at which the sign would be legible for a 90-year-old. Using the least squares regression line again as our summary of the linear dependence of the distances upon the drivers' ages, the agency predicts that 90-year-old drivers can see the sign at no more than 576 + (- 3 * 90) = 306 feet:

The scatterplot for Driver Age vs. Sign Legibility Distance. The scales of both axes have been enlarged so that the regression line has room on the right to be extended past where data exists. The regression line is negative, so it grows from the upper left to the lower right of the plot. Where the regression line is creating an estimate in between existing data, it is red. Beyond that, where there are no data points, the line is green. This area is x

82. The equation of the regression line is Distance = 576 - 3 * Age" height="274" loading="lazy" src="http://phhp-faculty-cantrell.sites.m...2-linear16.gif" title="The scatterplot for Driver Age vs. Sign Legibility Distance. The scales of both axes have been enlarged so that the regression line has room on the right to be extended past where data exists. The regression line is negative, so it grows from the upper left to the lower right of the plot. Where the regression line is creating an estimate in between existing data, it is red. Beyond that, where there are no data points, the line is green. This area is x>82. The equation of the regression line is Distance = 576 - 3 * Age" width="405">

(The green segment of the line is the region of ages beyond 82, the age of the oldest individual in the study.)

Question: Is our prediction for 90-year-old drivers reliable?

Answer: Our original age data ranged from 18 (youngest driver) to 82 (oldest driver), and our regression line is therefore a summary of the linear relationship **in that age range only.** When we plug the value 90 into the regression line equation, we are assuming that the same linear relationship extends beyond the range of our age data (18-82) into the green segment. **There is no justification for such an assumption.** It might be the case that the vision of drivers older than 82 falls off more rapidly than it does for younger drivers. (i.e., the slope changes from -3 to something more negative). Our prediction for age = 90 is therefore **not reliable.**

In General

Prediction for ranges of the explanatory variable that are not in the data is called **extrapolation**. Since there is no way of knowing whether a relationship holds beyond the range of the explanatory variable in the data, extrapolation is not reliable, and should be avoided. In our example, like most others, extrapolation can lead to very poor or illogical predictions.

> **Interactive Applets:** Linear Regression

> **Learn By Doing:** Linear Regression (Software)

Let's Summarize

- A special case of the relationship between two quantitative variables is the **linear** relationship. In this case, a straight line simply and adequately summarizes the relationship.
- When the scatterplot displays a linear relationship, we supplement it with the **correlation coefficient (r)**, which measures the **strength** and direction of a linear relationship between two quantitative variables. The correlation ranges between -1 and 1. Values near -1 indicate a strong negative linear relationship, values near 0 indicate a weak linear relationship, and values near 1 indicate a strong positive linear relationship.
- The correlation is only an appropriate numerical measure for linear relationships, and is sensitive to outliers. Therefore, the correlation should only be used as a supplement to a scatterplot (after we look at the data).
- The most commonly used criterion for finding a line that summarizes the pattern of a linear relationship is "least squares." The **least squares regression line** has the smallest sum of squared vertical deviations of the data points from the line.
- The **slope** of the least squares regression line can be interpreted as the estimated (or predicted) **change in the mean (or average) value of the response variable when the explanatory variable increases by 1 unit.**
- The **intercept** of the least squares regression line is the average value of the response variable when the explanatory variable is zero. Thus, this is only of interest if it makes sense for the explanatory variable to be zero AND we have observed data in that range (explanatory variable around zero) in our sample.
- The least squares regression line predicts the value of the response variable for a given value of the explanatory variable. **Extrapolation** is prediction of values of the explanatory variable that fall outside the range of the data. Since there is no way of knowing whether a relationship holds beyond the range of the explanatory variable in the data, extrapolation is not reliable, and should be avoided.

Causation

CO-1: Describe the roles biostatistics serves in the discipline of public health.

Video

Video: Causation (8:45)

Introduction

Learning Objectives

LO 1.6: Recognize the distinction between association and causation.

Learning Objectives

LO 1.7: Identify potential lurking variables for explaining an observed relationship.

So far we have discussed different ways in which data can be used to explore the relationship (or association) between two variables. To frame our discussion we followed the role-type classification table:

		Response	
		Categorical	Quantitative
Explanatory	Categorical	✓C → C	✓C → Q
	Quantitative	✗**Q → C**	✓Q → Q

We have now completed learning how to explore the relationship in cases C → Q, C → C, and Q → Q. (As noted before, case Q → C will not be discussed in this course.)

When we explore the relationship between two variables, there is often a temptation to conclude from the observed relationship that changes in the explanatory variable **cause** changes in the response variable. In other words, you might be tempted to interpret the observed association as causation.

The purpose of this part of the course is to convince you that this kind of interpretation is often **wrong!** The motto of this section is one of the most fundamental principles of this course:

WORDS TO LIVE BY: Statistical analysis alone will never prove causation!

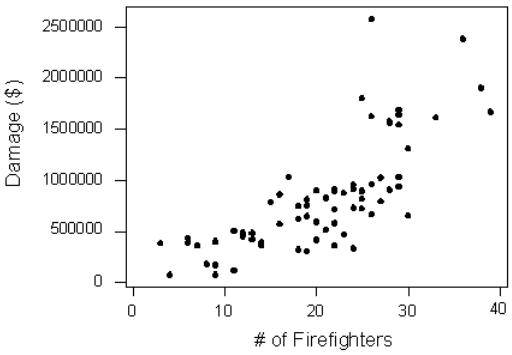
PRINCIPLE: Association does not imply causation!

Outside Reading: Cause & Effect (≈ 1700 words)

Let's start by looking at the following example:

EXAMPLE: Fire Damage

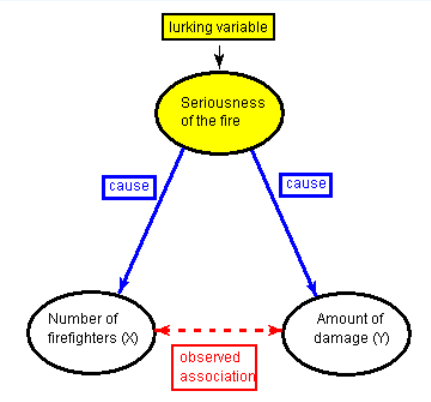
The scatterplot below illustrates how the number of firefighters sent to fires (X) is related to the amount of damage caused by fires (Y) in a certain city.



The scatterplot clearly displays a fairly strong (slightly curved) **positive** relationship between the two variables. Would it, then, be reasonable to conclude that sending more firefighters to a fire causes more damage, or that the city should send fewer firefighters to a fire, in order to decrease the amount of damage done by the fire? Of course not! So what is going on here?

There is a **third variable in the background** — the seriousness of the fire — that is responsible for the observed relationship. More serious fires require more firefighters, and also cause more damage.

The following figure will help you visualize this situation:



Here, the seriousness of the fire is a **lurking variable.** A **lurking variable** is a variable that is not among the explanatory or response variables in a study, but could substantially affect your interpretation of the relationship among those variables.

Here we have the following three relationships:

- Damage increases with the number of firefighters
- Number of firefighters increases with severity of fire
- Damage increases with the severity of fire
- Thus the increase in damage with the number of firefighters may be partially or fully explained by severity of fire.

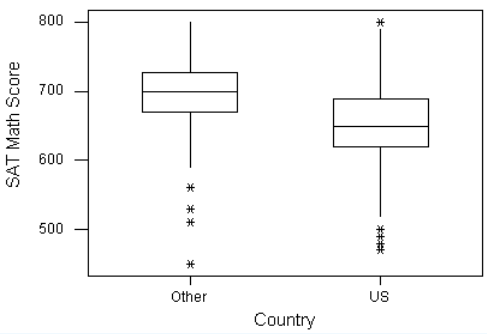
In particular, as in our example, the lurking variable might have an effect on ***both*** the explanatory and the response variables. This common effect creates the observed association between the explanatory and response variables, even though there is no causal link between them. This possibility, that there might be a lurking variable (which we might not be thinking about) that is responsible for the observed relationship leads to our principle:

PRINCIPLE: Association does not imply causation!

The next example will illustrate another way in which a lurking variable might interfere and prevent us from reaching any causal conclusions.

EXAMPLE: SAT Test

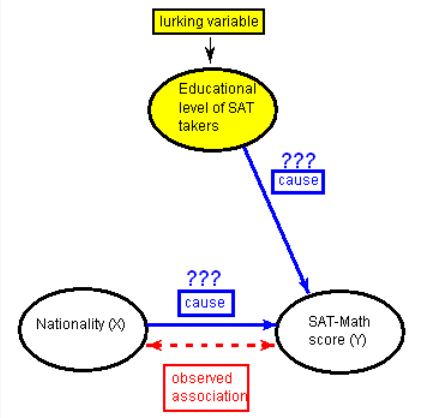
For U.S. colleges and universities, a standard entrance examination is the SAT test. The side-by-side boxplots below provide evidence of a relationship between the student's country of origin (the United States or another country) and the student's SAT Math score.



The distribution of international students' scores is higher than that of U.S. students. The international students' median score (about 700) exceeds the third quartile of U.S. students' scores. Can we conclude that the country of origin is the **cause** of the difference in SAT Math scores, and that students in the United States are weaker at math than students in other countries?

No, not necessarily. While it **might** be true that U.S. students differ in math ability from other students — i.e. due to differences in educational systems — we can't conclude that a student's country of origin is the cause of the disparity. One important **lurking variable** that might explain the observed relationship is the educational level of the two populations taking the SAT Math test. In the United States, the SAT is a standard test, and therefore a broad cross-section of all U.S. students (in terms of educational level) take this test. Among all international students, on the other hand, only those who plan on coming to the U.S. to study, which is usually a more selected subgroup, take the test.

The following figure will help you visualize this explanation:



Here, the explanatory variable (X) **may** have a causal relationship with the response variable (Y), but the lurking variable might be a contributing factor as well, which makes it very hard to isolate the effect of the explanatory variable and prove that it has a causal link with the response variable. In this case, we say that the lurking variable is **confounded** with the explanatory variable, since their effects on the response variable cannot be distinguished from each other.

Note that in each of the above two examples, the lurking variable interacts differently with the variables studied. In example 1, the lurking variable has an effect on both the explanatory and the response variables, creating the illusion that there is a causal link between them. In example two, the lurking variable is confounded with the explanatory variable, making it hard to assess the isolated effect of the explanatory variable on the response variable.

The distinction between these two types of interactions is not as important as the fact that in either case, the observed association can be at least partially explained by the lurking variable. The most important message from these two examples is therefore: **An observed association between two variables is not enough evidence that there is a causal relationship between them.**

In other words …

PRINCIPLE: Association does not imply causation!

Learn By Doing: Causation

Simpson's Paradox

Learning Objectives

LO 1.8: Recognize and explain the phenomenon of Simpson's Paradox as it relates to interpreting the relationship between two variables.

So far, we have:

- discussed what lurking variables are,
- demonstrated different ways in which the lurking variables can interact with the two studied variables, and
- understood that the existence of a possible lurking variable is the main reason why we say that association does not imply causation.

As you recall, a lurking variable, by definition, is a variable that was not included in the study, but could have a substantial effect on our understanding of the relationship between the two studied variables.

What if we **did** include a lurking variable in our study? What kind of effect could that have on our understanding of the relationship? These are the questions we are going to discuss next.

Let's start with an example:

EXAMPLE: Hospital Death Rates

Background: A government study collected data on the death rates in nearly 6,000 hospitals in the United States. These results were then challenged by researchers, who said that the federal analyses failed to take into account the variation among hospitals in the severity of patients' illnesses when they were hospitalized. As a result, said the researchers, some hospitals were treated unfairly in the findings, which named hospitals with higher-than-expected death rates. What the researchers meant is that when the federal government explored the relationship between the two variables — hospital and death rate — **it also should have included in the study (or taken into account) the lurking variable — severity of illness.**

We will use a simplified version of this study to illustrate the researchers' claim, and see what the possible effect could be of including a lurking variable in a study. (Reference: Moore and McCabe (2003). *Introduction to the Practice of Statistics*.)

Consider the following two-way table, which summarizes the data about the status of patients who were admitted to two hospitals in a certain city (Hospital A and Hospital B). Note that since the purpose of the study is to examine whether there is a "hospital effect" on patients' status, "Hospital is the explanatory variable, and "Patient's Status" is the response variable.

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

When we supplement the two-way table with the conditional percents within each hospital:

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**3%**	97%	100%
	Hospital B	**2%**	98%	100%

we find that Hospital A has a higher death rate (3%) than Hospital B (2%). Should we jump to the conclusion that a sick patient admitted to Hospital A is 50% more likely to die than if he/she were admitted to Hospital B? **Not so fast …**

Maybe Hospital A gets most of the severe cases, and that explains why it has a higher death rate. In order to explore this, we need to **include (or account for) the lurking variable "severity of illness" in our analysis.** To do this, we go back to the two-way table and split it up to look separately at patients who are severely ill, and patients who are not.

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

Accounting for the lurking variable: **"severity of illness"**

Patients severely ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	57	1443	1500
	Hospital B	8	192	200
	Total	65	1635	1700

Patients *not* severely ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	6	594	600
	Hospital B	8	592	600
	Total	14	1186	1200

As we can see, Hospital A **did** admit many more severely ill patients than Hospital B (1,500 vs. 200). In fact, from the way the totals were split, we see that in Hospital A, severely ill patients were a much higher proportion of the patients — 1,500 out of a total of 2,100 patients. In contrast, only 200 out of 800 patients at Hospital B were severely ill. To better see the effect of including the lurking variable, we need to supplement each of the two new two-way tables with its conditional percentages:

Patients severely ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**3.8%**	96.2%	100%
	Hospital B	**4.0%**	96.0%	100%

Patients *not* severely ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**1.0%**	99.0%	100%
	Hospital B	**1.3%**	98.7%	100%

Note that despite our earlier finding that overall Hospital A has a higher death rate (3% vs. 2%), when we take into account the lurking variable, we find that actually it is Hospital B that has the higher death rate both among the severely ill patients (4% vs. 3.8%) and among the not severely ill patients (1.3% vs. 1%). **Thus, we see that adding a lurking variable can change the direction of an association.**

Here we have the following three relationships:

- A greater percentage of hospital A's patient's died compared to hospital B.
- Patient's who are severely ill are less likely to survive.
- Hospital A accepts more severely ill patients.
- In this case, after further careful analysis, we see that once we account for severity of illness, hospital A actually has a lower percentage of patient's who died than hospital B in both groups of patients!

Whenever including a lurking variable causes us to **rethink the direction** of an association, this is called **Simpson's paradox.**

The possibility that a lurking variable can have such a dramatic effect is another reason we must adhere to the principle:

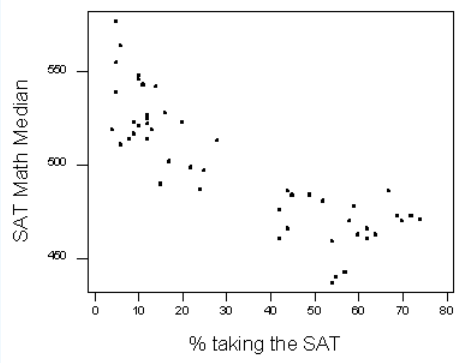
PRINCIPLE: Association does not imply causation!

A Final Example – Gaining a Deeper Understaing of the Relationship

It is **not** always the case that including a lurking variable makes us rethink the direction of the association. In the next example we will see how including a lurking variable just helps us gain a deeper understanding of the observed relationship.

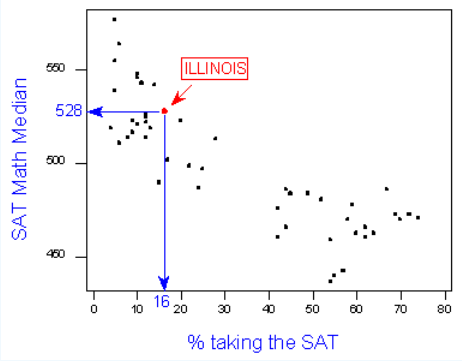
EXAMPLE: College Entrance Exams

As discussed earlier, in the United States, the SAT is a widely used college entrance examination, required by the most prestigious schools. In some states, a different college entrance examination is prevalent, the ACT.

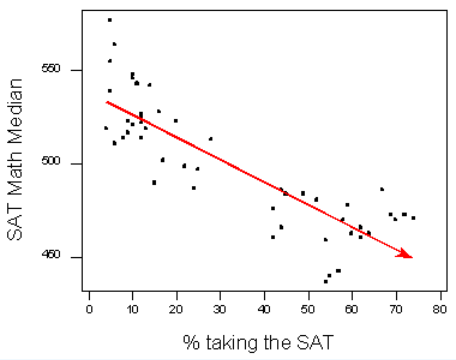


Note that:

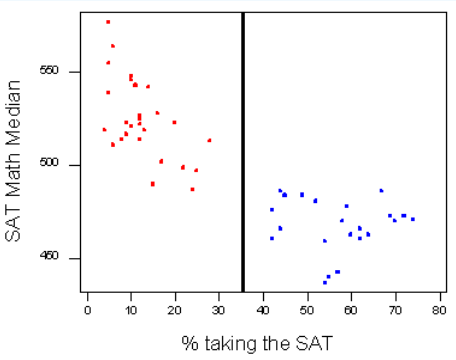
- the explanatory variable is the percentage taking the SAT,
- the response variable is the median SAT Math score, and
- each data point on the scatterplot represents one of the states, so for example, in Illinois, in the year these data were collected, 16% of the students took the SAT Math, and their median score was 528.



Notice that there is a negative relationship between the percentage of students who take the SAT in a state, and the median SAT Math score in that state. What could the explanation behind this negative trend be? Why might having more people take the test be associated with lower scores?

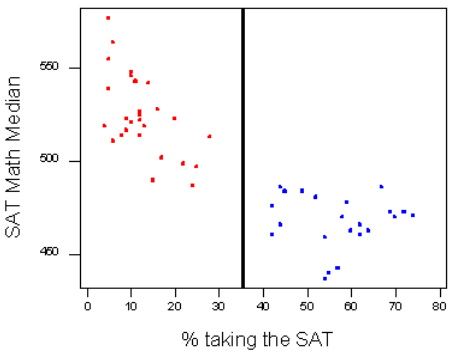


Note that another visible feature of the data is the presence of a gap in the middle of the scatterplot, which creates two distinct clusters in the data. This suggests that maybe there is a lurking variable that separates the states into these two clusters, and that including this lurking variable in the study (as we did, by creating this labeled scatterplot) will help us understand the negative trend.



It turns out that indeed, the clusters represent two groups of states:

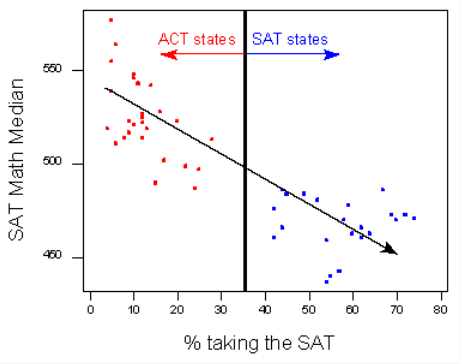
- The “blue group” on the right represents the states where the SAT is the test of choice for students and colleges.
- The “red group” on the left represents the states where the ACT college entrance examination is commonly used.



It makes sense then, that in the “ACT states” on the left, a smaller percentage of students take the SAT. Moreover, the students who do take the SAT in the ACT states are probably students who are applying to more prestigious national colleges, and therefore represent a more select group of students. This is the reason why we see high SAT Math scores in this group.

On the other hand, in the “SAT states” on the right, larger percentages of students take the test. These students represent a much broader cross-section of the population, and therefore we see lower (more average) SAT Math scores.

To summarize: In this case, including the lurking variable “ACT state” versus “SAT state” helped us better understand the observed negative relationship in our data.



Learn By Doing: Causation and Lurking Variables

Did I Get This?: Simpson's Paradox

The last two examples showed us that including a lurking variable in our exploration may:

- lead us to **rethink the direction** of an association (as in the Hospital/Death Rate example) or,
- help us to **gain a deeper understanding of the relationship** between variables (as in the SAT/ACT example).

Let's Summarize

- A **lurking variable** is a variable that was not included in your analysis, but that could substantially change your interpretation of the data if it were included.
- Because of the possibility of lurking variables, we adhere to the principle that ***association does not imply causation***.
- Including a lurking variable in our exploration may:
  - help us to **gain a deeper understanding** of the relationship between variables, or
  - lead us to **rethink the direction of an association (Simpson's Paradox)**
- Whenever including a lurking variable causes us to **rethink the direction of an association**, this is an instance of **Simpson's paradox**.

One Categorical Variable

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Video

One Categorical Variable (4:57)

Note

Note: These videos are listed for reference. If you would like to follow along in your first reading, then you will need to see the preceding tutorial videos. These videos are also linked in the programming assignments.

- All SAS tutorial videos
- All SPSS tutorial videos

Related SAS Tutorials

- 4A – (3:03) Frequency Distributions

Related SPSS Tutorials

- 4A – (7:00) Frequency Distributions

Distribution of One Categorical Variable

Learning Objectives

LO 4.3: Using appropriate numerical measures and/or visual displays, describe the distribution of a categorical variable in context.

What is your perception of your own body? Do you feel that you are overweight, underweight, or about right?

A random sample of 1,200 U.S. college students were asked this question as part of a larger survey. The following table shows part of the responses:

Student	Body Image
student 25	overweight
student 26	about right
student 27	underweight
student 28	about right
student 29	about right

Here is some information that would be interesting to get from these data:

- What percentage of the sampled students fall into each category?
- How are students divided across the three body image categories? Are they equally divided? If not, do the percentages follow some other kind of pattern?

There is no way that we can answer these questions by looking at the raw data, which are in the form of a long list of 1,200 responses, and thus not very useful.

Both of these questions will be easily answered once we summarize and look at the **distribution** of the variable Body Image (i.e., once we summarize how often each of the categories occurs).

Numerical Measures

In order to summarize the distribution of a **categorical** variable, we first create a table of the different values (categories) the variable takes, how many times each value occurs (count) and, more importantly, how often each value occurs (by converting the counts to percentages).

The result is often called a **Frequency Distribution** or **Frequency Table**.

> **Note**
>
> A **Frequency Distribution** or **Frequency Table** is the primary **set of numerical measures** for **one categorical variable**.
>
> - Consists of a **table** with **each category** along with the **count** and **percentage** for each category.
> - Provides a summary of the **distribution** for **one categorical variable**.

Here is the table for our example:

Category	Count	Percent
About right	855	(855/1200)*100 = 71.3%
Overweight	235	(235/1200)*100 = 19.6%
Underweight	110	(110/1200)*100 = 9.2%
Total	**n=1200**	***100%***

Comments:

1. If you add the percentages in the above table you will get a total of 100.1% (instead of the true value which is, of course, 100%).This can occur whenever rounding has taken place. You should be aware of this possibility when working with real data.If you add the ratios directly as fractions, you will always get exactly 1 (or 100%).

---

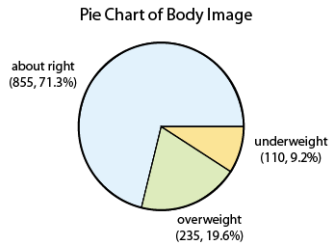
2. In general, although it might be “less confusing” if we recorded the full values above (71.25% instead of 71.3% and so on), we prefer not to display too many decimal places as this can distract from the conclusions we want to illustrate.We don’t want those who are reading our results to be overwhelmed or distracted by unneeded digits.

Visual or Graphical Displays

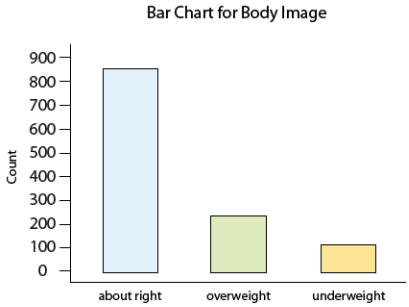
In order to visualize the numerical measures we’ve obtained, we need a graphical display.

> **Note**
>
> There are two simple **graphical displays** for **visualizing** the **distribution of one categorical variable**:
>
> - **Pie Charts**
> - **Bar Charts**

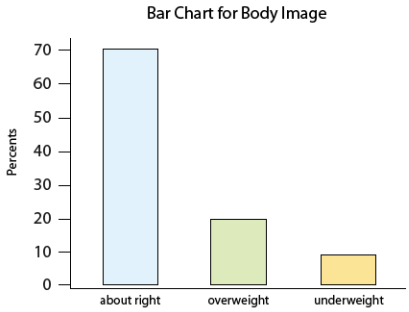
Pie Chart



Bar Chart



OR



Note that the pie chart and bar chart are visual representations of the information in the frequency table.

Study the bar charts above and then answer the following question.

Learn By Doing: Bar Charts

Now that we have summarized the distribution of values in the Body Image variable, let's go back and interpret the results in the context of the questions that we posed. Study the frequency table and graphs and answer the following questions.

Learn By Doing: Describe the Distribution of a Categorical Variable

Now that we've interpreted the results, there are some other interesting questions that arise:

- Can we reliably generalize our results to the entire population of interest and conclude that a similar distribution across body image categories exists among all U.S. college students? In particular, can we make such a generalization even though our sample consisted of only 1,200 students, which is a very small fraction of the entire population?
- If we had separated our sample by gender and looked at males and females separately, would we have found a similar distribution across body image categories?

These are the types of questions that we will deal with in future sections of the course.

Recall: Categorical variables take category or label values, and place an individual into one of several groups. Categorical variables are often further classified as either

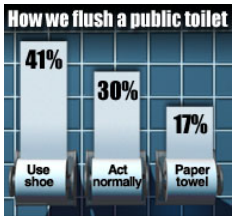
- **Nominal,** when there is no natural ordering among the categories. Common examples would be gender, eye color, or ethnicity.
- **Ordinal**, when there is a natural order among the categories, such as, ranking scales or letter grades. However, ordinal variables are categorical and do not provide precise measurements. Differences are not precisely meaningful, for example, if one student scores an A and another a B on an assignment, we cannot say precisely the difference in their scores, only that an A is larger than a B.

Note: For ordinal categorical variables, pie charts are seldom used since the information about the order can be lost in such a display. Be careful that bar charts for ordinal variables display the data in a reasonable order given the scenario.

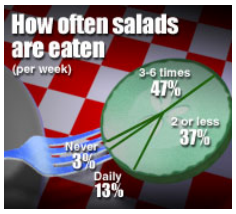
While both the pie chart and the bar chart help us visualize the distribution of a categorical variable, the pie chart emphasizes how the different categories relate to the whole, and the bar chart emphasizes how the different categories compare with each other.

Pictograms

A variation on the pie chart and bar chart that is very commonly used in the media is the pictogram. Here are two examples:

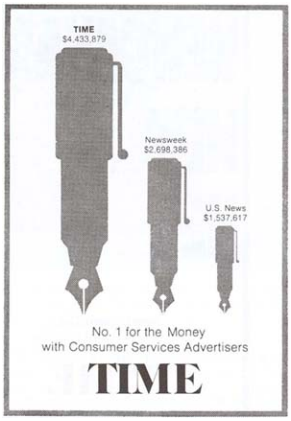


Source: USA Today Snapshots and the Impulse Research for Northern Confidential Bathroom survey



Source: Market Facts for the Association of Dressings and Sauces

Beware: Pictograms can be misleading. Consider the following pictogram:



This graph is aimed at advertisers deciding where to spend their budgets, and clearly suggests that Time magazine attracts by far the largest amount of advertising spending.

Are the differences really as dramatic as the graph suggests?

If we look carefully at the numbers above the pens, we find that advertisers spend in Time only $4,433,879 / $2,698,386 = 1.64 times more than in Newsweek, and only $4,433,879 / $1,537,617 = 2.88 times more than in U.S. News.

By looking at the pictogram, however, we get the impression that Time is much further ahead. Why?

In order to magnify the picture without distorting it, we must increase both its height and width. As a result, the area of Time's pen is 1.64 * 1.64 = 2.7 times larger than the Newsweek pen, and 2.88 * 2.88 = 8.3 times larger than the U.S. News pen. Our eyes capture the area of the pens rather than only the height, and so we are misled to think that Time is a bigger winner than it really is.

Learn By Doing: One Categorical Variable (College Student Survey)

Let's Summarize

The distribution of a categorical variable is summarized using:

- **Visual display:** pie chart or bar chart, supplemented by
- **Numerical measures:** frequency table of category counts and percentages.

A variation on pie charts and bar charts is the pictogram. Pictograms can be misleading, so make sure to use a critical approach when interpreting the information the pictogram is trying to convey.

One Quantitative Variable: Introduction

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Video

Video: One Quantitative Variable (4:16)

Note

Related SAS Tutorials

- 5A – (3:01) Numeric Measures using PROC MEANS
- 5B – (4:05) Creating Histograms and Boxplots using SGPLOT
- 5C – (5:41) Creating QQ-Plots and other plots using UNIVARIATE

Related SPSS Tutorials

- 5A – (8:00) Numeric Measures using EXPLORE
- 5B – (2:29) Creating Histograms and Boxplots
- 5C – (2:31) Creating QQ-Plots and PP-Plots

Distribution of One Quantitative Variable

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

In the previous section, we explored the distribution of a categorical variable using graphs (pie chart, bar chart) supplemented by numerical measures (percent of observations in each category).

In this section, we will explore the data collected from a **quantitative** variable, and learn how to describe and summarize the important features of its distribution.

We will learn how to display the **distribution** using **graphs** and discuss a variety of **numerical measures**.

An introduction to each of these topics follows.

Graphs

To display data from one quantitative variable graphically, we can use either a **histogram** or **boxplot**.

We will also present several "by-hand" displays such as the **stemplot** and **dotplot** (although we will not rely on these in this course).

Numerical Measures

The overall pattern of the **distribution** of a quantitative variable is described by its **shape**, **center**, and **spread**.

By inspecting the histogram or boxplot, we can describe the shape of the distribution, but we can only get a rough estimate for the center and spread.

A description of the distribution of a quantitative variable must include, in addition to the **graphical display**, a more precise **numerical description** of the center and spread of the distribution.

In this section we will learn:

- how to display the **distribution of one quantitative variable** using various graphs;
- how to quantify the **center** and **spread** of the **distribution of one quantitative variable** with various numerical measures;

- some of the **properties** of those **numerical measures**;
- how to choose the **appropriate numerical measures** of **center** and **spread** to supplement the graph(s); and
- how to identify potential outliers in the **distribution of one quantitative variable**
- We will also discuss a few **measures of position** (also called **measures of location**). These measures
  - allow us to quantify where a particular value is relative to the **distribution** of all values
  - do provide information about the distribution itself
  - also use the information **about the distribution** to learn more about an **INDIVIDUAL**

We will present the material in a logical sequence which builds in difficulty, intermingling discussion of visual displays and numerical measures as we proceed.

Before reading further, try this interactive applet which will give you a preview of some of the topics we will be learning about in this section on exploratory data analysis for one quantitative variable.

Interactive Applet: Analyze One Quantitative Variable with this One-Variable Statistical Calculator

Histograms & Stemplots

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Video

Video: Histograms and Stemplots (5:03)

Note

Related SAS Tutorials

- 5B – (4:05) Creating Histograms and Boxplots using SGPLOT

Related SPSS Tutorials

- 5B – (2:29) Creating Histograms and Boxplots

Histograms

Learning Objectives

LO 4.5: Explain the process of creating a histogram.

The idea is to break the range of values into intervals and count how many observations fall into each interval.

EXAMPLE: Exam Grades

Here are the exam grades of 15 students:

88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

We first need to break the range of values into intervals (also called “bins” or “classes”).

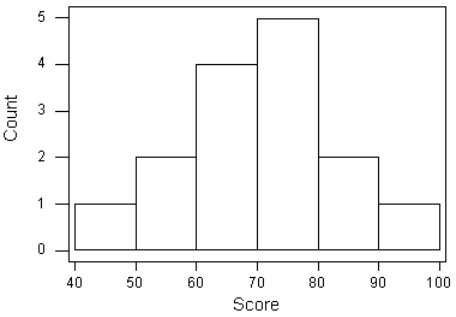
In this case, since our dataset consists of exam scores, it will make sense to choose intervals that typically correspond to the range of a letter grade, 10 points wide: [40,50), [50, 60), … [90, 100).

By counting how many of the 15 observations fall in each of the intervals, we get the following table:

Score	Count
[40-50)	1
[50-60)	2
[60-70)	4
[70-80)	5
[80-90)	2
[90-100)	1

Note: The observation 60 was counted in the 60-70 interval. See comment 1 below.

To construct the histogram from this table we plot the intervals on the X-axis, and show the number of observations in each interval (frequency of the interval) on the Y-axis, which is represented by the height of a rectangle located above the interval:



The previous table can also be turned into a relative frequency table using the following steps:

- Add a row on the bottom and include the total number of observations in the dataset that are represented in the table.
- Add a column, at the end of the table, and calculate the relative frequency for each interval, by dividing the number of observations in each row by the total number of observations.

These two steps are illustrated in red in the following frequency distribution table:

Score	Count (also called Frequency)	Relative Frequency
[40–50)	1	0.07
[50–60)	2	0.13
[60–70)	4	0.27
[70–80)	5	0.33
[80–90)	2	0.13
[90–100]	1	0.07
Total	15	

Step 1: Add a row at bottom of table. Put in total number of observations in the data set.

In this example, there are 15 (1+2+4+5+2+1= 15) total observations.

Step 2: Add a column to right side of table. Determine the relative frequencies of each interval by dividing the interval count (or frequency) by the total number of observations.

For example, to determine the relative frequency of scores in the [40-50) interval, divide the count (or frequency) by the total number of observations: 1 /15 = .07.

The relative frequency for the [50-60) interval is: 2/15 = .13.

Continue until all of the relative frequencies have been calculated.

To convert each relative frequency into a percentage, multiply it by 100. For example, the percentage of scores for the [40-50) interval would be .07*100= 7, which is 7%.

It is also possible to determine the number of scores for an interval, if you have the total number of observations and the relative frequency for that interval.

- For instance, suppose there are 15 scores (or observations) in a set of data and the relative frequency for an interval is 0.13.
- To determine the number of scores in that interval, multiplying the total number of observations by the relative frequency and round up to the next whole number: 15*.13 = 1.95, which rounds up to 2 observations.

A relative frequency table, like the one above, can be used to determine the frequency of scores occurring at or across intervals.

Here are some examples, using this frequency table:

What is the percentage of exam scores that were 70 and up to, but not including, 80?

- To determine the answer, we look at the relative frequency associated with the [70-80) interval.
- The relative frequency is 0.33; to convert to percentage, multiply by 100 (0.33*100= 33) or 33%.

What is the percentage of exam scores that are at least 70? To determine the answer, we need to:

- Add together the relative frequencies for the intervals that have scores of at least 70 or above.
- Thus, would need to add together the relative frequencies from [70-80), [80-90), and [90-100] = 0.33 + 0.13 + 0.07 = 0.53.
- To get the percentage, need to multiple the calculated relative frequency by 100.
- In this case, it would be 0.53*100 = 53 or 53%.

Study the histogram again and table and answer the following question.

Learn By Doing: Histograms

Comments:

- It is very important that each observation be counted only in one interval. For the most part, it is clear which interval an observation falls in. However, in our example, we needed to decide whether to include 60 in the interval 50-60, or the interval 60-70, and we chose to count it in the latter.
  - In fact, this decision is captured by the way we wrote the intervals. If you'll scroll up and look at the table, you'll see that we wrote the intervals in a peculiar way: [40-50), [50,60), [60,70) etc.
  - The square bracket means "including" and the parenthesis means "not including". For example, [50,60) is the interval from 50 to 60, including 50 and not including 60; [60,70) is the interval from 60 to 70, including 60, and not including 70, etc.
  - It really does not matter how you decide to set up your intervals, as long as you are consistent.
  - When you look at a histogram such as the one above it is important to know that values falling on the border are only counted in one interval, even if you do not know which way this was done for a particular graph.

- When data are displayed in a histogram, some information is lost. Note that by looking at the histogram
  - we ***can*** answer: "How many students scored 70 or above?" (5+2+1=8)
  - But we ***cannot*** answer: "What was the lowest score?" All we can say is that the lowest score is somewhere between 40 and 50.

- Obviously, we could have chosen to break the data into intervals differently — for example: [45, 50), [50, 55), [55, 60) etc.

To see how our choice of bins or intervals affects a histogram, you can use the applet linked below that let you change the intervals dynamically.

(OPTIONAL) Interactive Applet: Histograms

Many Students Wonder: Histograms

Question*:* How do I know what interval width to choose?

Answer: There are many valid choices for interval widths and starting points. There are a few rules of thumb used by software packages to find optimal values. In this course, we will rely on a statistical package to produce the histogram for us, and we will focus instead on describing and summarizing the distribution as it appears from the histogram.

The following exercises provide more practice working with histograms created from a single quantitative variable.

Did I Get This?: Histograms

Stemplot (Stem and Leaf Plot)

Learning Objectives

LO 4.6: Explain the process of creating a stemplot.

The **stemplot** (also called stem and leaf plot) is another graphical display of the distribution of quantitative variable.

Note

To create a **stemplot, t**he idea is to separate each data point into a stem and leaf, as follows:

- The leaf is the right-most digit.
- The stem is everything except the right-most digit.
- So, if the data point is 34, then 3 is the stem and 4 is the leaf.
- If the data point is 3.41, then 3.4 is the stem and 1 is the leaf.
- Note: For this to work, ALL data points should be rounded to the same number of decimal places.

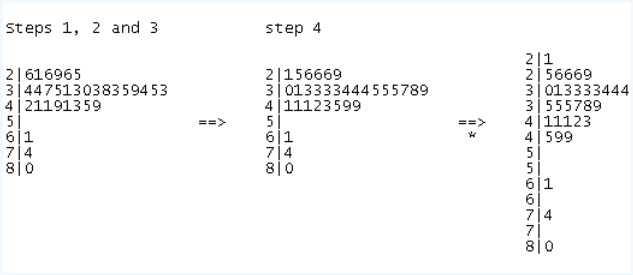
EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example (Link to the Best Actress Oscar Winners data).

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33

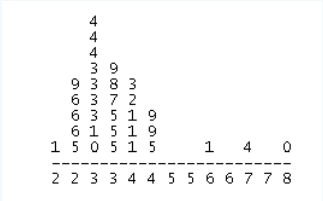
To make a stemplot:

- Separate each observation into a stem and a leaf.
- Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
- Go through the data points, and write each leaf in the row to the right of its stem.
- Rearrange the leaves in an increasing order.



* When some of the stems hold a large number of leaves, we can split each stem into two: one holding the leaves 0-4, and the other holding the leaves 5-9. A statistical software package will often do the splitting for you, when appropriate.

Note that when rotated 90 degrees counterclockwise, the stemplot visually resembles a histogram:



The stemplot has additional unique features:

- It preserves the original data.
- It sorts the data (which will become very useful in the next section).

You will not need to create these plots by hand but you may need to be able to discuss the information they contain.

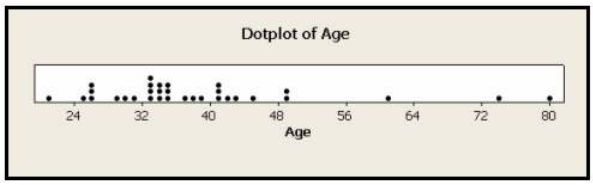
To see more stemplots, use the interactive applet we introduced earlier.

In particular, notice how the raw data are rounded and look at the stemplot with and without split stems.

Interactive Applet: Analyze One Quantitative Variable with this One-Variable Statistical Calculator

Comments: ABOUT DOTPLOTS

- There is another type of display that we can use to summarize a quantitative variable graphically — the dotplot.
- The dotplot, like the stemplot, shows each observation, but displays it with a dot rather than with its actual value.
- We will not use these in this course but you may see them occasionally in practice and they are relatively easy to create by-hand.
- Here is the dotplot for the ages of Best Actress Oscar winners.



Many Students Wonder: Graphs

Question: How do we know which graph to use: the histogram, stemplot, or dotplot?

Answer Since for the most part we are not going to deal with very small data sets in this course, we will generally display the distribution of a quantitative variable using a histogram generated by a statistical software package.

Let's Summarize

- The histogram is a graphical display of the distribution of a quantitative variable. It plots the number (count) of observations that fall in intervals of values.
- The stemplot is a simple, but useful visual display of a quantitative variable. Its principal virtues are:
  - Easy and quick to construct for small, simple datasets.
  - Retains the actual data.
  - Sorts (ranks) the data.

Describing Distributions

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Video

Video: Describing Distributions (2 videos, 7:38 total)

Note

Related SAS Tutorials

- 5A – (3:01) Numeric Measures using PROC MEANS
- 5B – (4:05) Creating Histograms and Boxplots using SGPLOT
- 5C – (5:41) Creating QQ-Plots and other plots using UNIVARIATE

Related SPSS Tutorials

- 5A – (8:00) Numeric Measures using EXPLORE
- 5B – (2:29) Creating Histograms and Boxplots
- 5C – (2:31) Creating QQ-Plots and PP-Plots

Features of Distributions of Quantitative Variables

Learning Objectives

LO 4.7: Define and describe the features of the distribution of one quantitative variable (shape, center, spread, outliers).

Once the distribution has been displayed graphically, we can describe the overall pattern of the distribution and mention any striking deviations from that pattern.

Note

More specifically, we should consider the following features of the Distribution for One Quantitative Variable:

- Shape
- Center
- Spread

} overall pattern

- Outliers → deviations from the pattern

Shape

When describing the shape of a distribution, we should consider:

- **Symmetry/skewness** of the distribution.
- **Peakedness (modality)** — the number of peaks (modes) the distribution has.

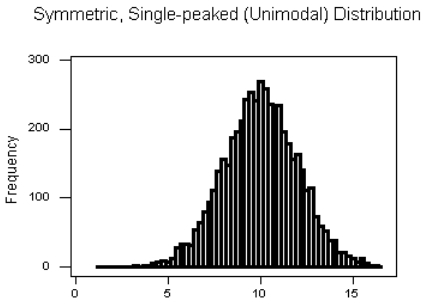
We distinguish between:

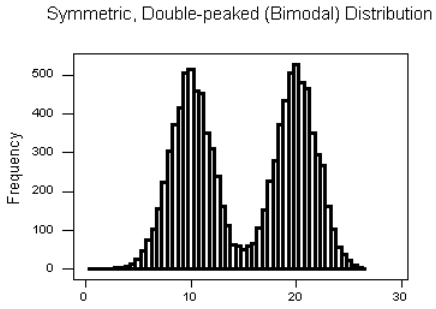
Symmetric Distributions

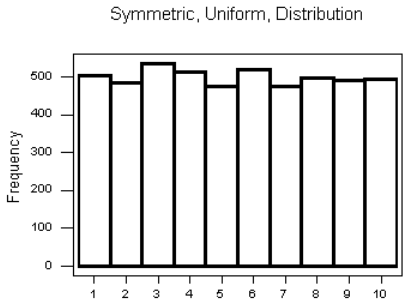
Note

A distribution is called **symmetric** if, as in the histograms above, the distribution forms an approximate mirror image with respect to the center of the distribution.

The center of the distribution is easy to locate and both tails of the distribution are the approximately the same length.



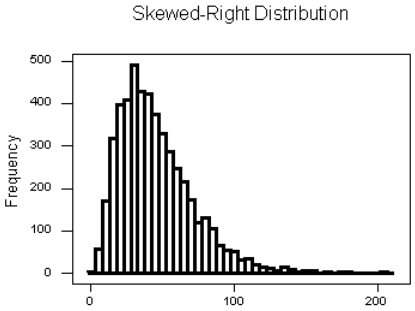




Note that all three distributions are symmetric, but are different in their **modality** (peakedness).

- The first distribution is **unimodal** — it has one mode (roughly at 10) around which the observations are concentrated.
- The second distribution is **bimodal** — it has two modes (roughly at 10 and 20) around which the observations are concentrated.
- The third distribution is kind of flat, or **uniform**. The distribution has no modes, or no value around which the observations are concentrated. Rather, we see that the observations are roughly uniformly distributed among the different values.

Skewed Right Distributions

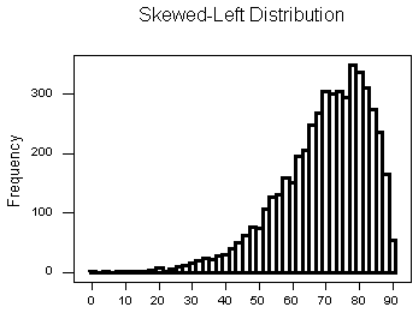


> A distribution is called **skewed right** if, as in the histogram above, the right tail (larger values) is much longer than the left tail (small values).

Note that in a skewed right distribution, the bulk of the observations are small/medium, with a few observations that are much larger than the rest.

- An example of a real-life variable that has a skewed right distribution is salary. Most people earn in the low/medium range of salaries, with a few exceptions (CEOs, professional athletes etc.) that are distributed along a large range (long “tail”) of higher values.

Skewed Left Distributions



A distribution is called **skewed left** if, as in the histogram above, the left tail (smaller values) is much longer than the right tail (larger values).

Note that in a skewed left distribution, the bulk of the observations are medium/large, with a few observations that are much smaller than the rest.

- An example of a real life variable that has a skewed left distribution is age of death from natural causes (heart disease, cancer etc.). Most such deaths happen at older ages, with fewer cases happening at younger ages.

Comments:

1. Distributions with more than two peaks are generally called **multimodal**.

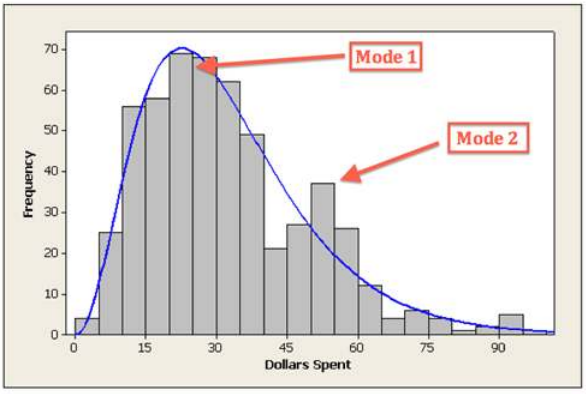
---

2. Bimodal or multimodal distributions can be evidence that two distinct groups are represented.

---

3. Unimodal, Bimodal, and multimodal distributions may or may not be symmetric.

Here is an example. A medium size neighborhood 24-hour convenience store collected data from 537 customers on the amount of money spent in a single visit to the store. The following histogram displays the data.



Note that the overall shape of the distribution is skewed to the right with a clear mode around $25. In addition, it has another (smaller) “peak” (mode) around $50-55.

The majority of the customers spend around $25 but there is a cluster of customers who enter the store and spend around $50-55.

Center

The **center** of the distribution is often used to represent a typical value.

One way to define the center is as the value that divides the distribution so that approximately half the observations take smaller values, and approximately half the observations take larger values.

Another common way to measure the center of a distribution is to use the average value.

From looking at the histogram we can get only a rough estimate for the center of the distribution. More exact ways of finding measures of center will be discussed in the next section.

Spread

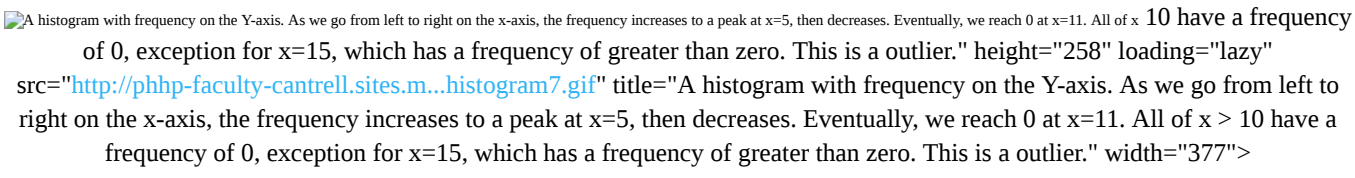
> One way to measure the **spread** (also called **variability** or **variation**) of the distribution is to use the approximate range covered by the data.

From looking at the histogram, we can approximate the smallest observation (**min**), and the largest observation (**max**), and thus approximate the **range**. (More exact ways of finding measures of spread will be discussed soon.)

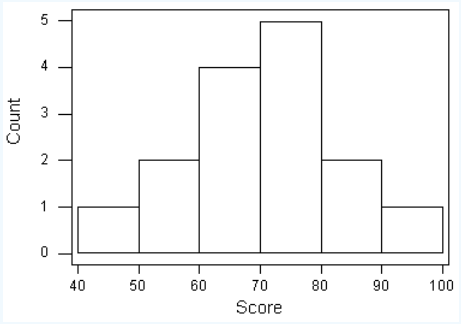
Outliers

> **Outliers** are observations that fall outside the overall pattern.

For example, the following histogram represents a distribution with a highly probable outlier:



EXAMPLE: Exam Grades



As you can see from the histogram, the grades distribution is roughly **symmetric** and **unimodal** with **no outliers**.

The **center** of the grades distribution is roughly **70** (7 students scored below 70, and 8 students scored above 70).

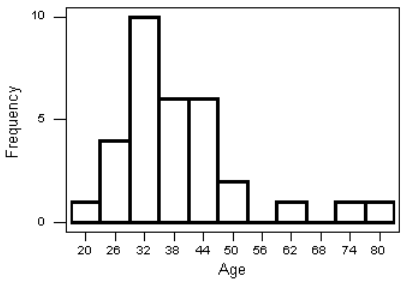
approximate min:	45 (the middle of the lowest interval of scores)
approximate max:	95 (the middle of the highest interval of scores)
approximate range:	95-45=50

Let's look at a new example.

EXAMPLE: Best Actress Oscar Winners

To provide an example of a histogram applied to actual data, we will look at the ages of Best Actress Oscar winners from 1970 to 2001

The histogram for the data is shown below. (Link to the Best Actress Oscar Winners data).



We will now summarize the main features of the distribution of ages as it appears from the histogram:

Shape: The distribution of ages is skewed right. We have a concentration of data among the younger ages and a long tail to the right. The vast majority of the "best actress" awards are given to young actresses, with very few awards given to actresses who are older.

Center: The data seem to be centered around 35 or 36 years old. Note that this implies that roughly half the awards are given to actresses who are less than 35 years old.

Spread: The data range from about 20 to about 80, so the approximate range equals 80 – 20 = 60.

Outliers: There seem to be two probable outliers to the far right and possibly a third around 62 years old.

You can see how informative it is to know "what to look at" in a histogram.

Learn By Doing: Shapes of Distributions (Best Actor Oscar Winners)

The following exercises provide more practice with shapes of distributions for one quantitative variable.

Did I Get This?: Shapes of Distributions

Did I Get This?: Shapes of Distributions Part 2

Let's Summarize

- When examining the distribution of a quantitative variable, one should describe the overall pattern of the data (shape, center, spread), and any deviations from the pattern (outliers).
- When describing the shape of a distribution, one should consider:
  - Symmetry/skewness of the distribution
  - Peakedness (modality) — the number of peaks (modes) the distribution has.
  - Not all distributions have a simple, recognizable shape.
- Outliers are data points that fall outside the overall pattern of the distribution and need further research before continuing the analysis.
- It is always important to interpret what the features of the distribution mean in the context of the data.

Measures of Center

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Learning Objectives

LO 4.7: Define and describe the features of the distribution of one quantitative variable (shape, center, spread, outliers).

Video

Video: Measures of Center (2 videos, 6:09 total)

Note

Related SAS Tutorials

- 5A – (3:01) Numeric Measures using PROC MEANS

Related SPSS Tutorials

- 5A – (8:00) Numeric Measures using EXPLORE

Introduction

Intuitively speaking, a numerical measure of center describes a "typical value" of the distribution.

The two main numerical measures for the center of a distribution are the **mean** and the **median**.

In this unit on Exploratory Data Analysis, we will be calculating these results based upon a **sample** and so we will often emphasize that the values calculated are the **sample mean** and **sample median**.

Each one of these measures is based on a completely different idea of describing the center of a distribution.

We will first present each one of the measures, and then compare their properties.

Mean

Learning Objectives

LO 4.8: Define and calculate the sample mean of a quantitative variable.

EXAMPLE

The **mean** is the **average** of a set of observations (i.e., the sum of the observations divided by the number of observations).

The **mean** is the **average** of a set of observations

- The sum of the observations divided by the number of observations).
- If the n observations are written as

$$x_1, x_2, \cdots, x_n$$

- their mean can be written mathematically as:their mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

We read the symbol as "x-bar." The bar notation is commonly used to represent the **sample mean**, i.e. the mean of the sample.

Using any appropriate letter to represent the variable (x, y, etc.), we can indicate the sample mean of this variable by adding a bar over the variable notation.

EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example (Link to the Best Actress Oscar Winners data).

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33

The mean age of the 32 actresses is:

$$\bar{x} = \frac{34+34+26+\ldots+35+33}{32} = \frac{1233}{32} = 38.5$$

We add all of the ages to get **1233** and **divide by** the number of ages which was **32** to get **38.5.**

We denote this result as **x-bar** and called the **sample mean**.

Note that the sample mean gives a measure of center which is higher than our approximation of the center from looking at the histogram (which was 35). The reason for this will be clear soon.

EXAMPLE: World Cup Soccer

Often we have large sets of data and use a frequency table to display the data more efficiently.

Data were collected from the last three World Cup soccer tournaments. A total of 192 games were played. The table below lists the number of goals scored per game (not including any goals scored in shootouts).

Total # Goals/Game	Frequency
0	17
1	45
2	51
3	37
4	25
5	11
6	3
7	2
8	1

To find the mean number of goals scored per game, we would need to **find the sum of all 192 numbers, and then divide that sum by 192.**

Rather than add 192 numbers, we use the fact that the same numbers appear many times. For example, the number 0 appears 17 times, the number 1 appears 45 times, the number 2 appears 51 times, etc.

If we add up 17 zeros, we get 0. If we add up 45 ones, we get 45. If we add up 51 twos, we get 102. Repeated addition is multiplication.

Thus, the **sum of the 192 numbers**

= 0(17) + 1(45) + 2(51) + 3(37) + 4(25) + 5(11) + 6(3) + 7(2) + 8(1) = 453.

The **sample mean** is then **453 / 192 = 2.359**.

Note that, in this example, the values of 1, 2, and 3 are the most common and our average falls in this range representing the bulk of the data.

Did I Get This?: Mean

Median

Learning Objectives

LO 4.9: Define and calculate the sample median of a quantitative variable.

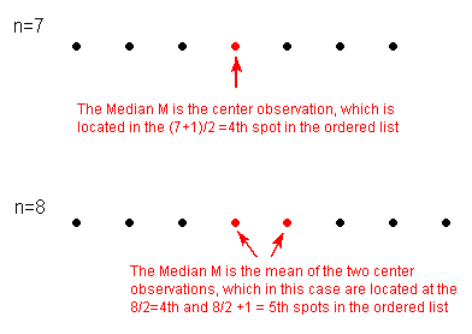
The **median** M is the midpoint of the distribution. It is the number such that half of the observations fall above, and half fall below.

To find the median:

- Order the data from smallest to largest.
- Consider whether n, the number of observations, is even or odd.
  - If n is **odd**, the median M is the center observation in the ordered list. This observation is the one "sitting" in the **(n + 1) / 2** spot in the ordered list.
  - If n is **even**, the median M is the **mean** of the **two center observations** in the ordered list. These two observations are the ones "sitting" in the **(n / 2)** and **(n / 2) + 1** spots in the ordered list.

EXAMPLE: Median(1)

For a simple visualization of the location of the median, consider the following two simple cases of n = 7 and n = 8 ordered observations, with each observation represented by a solid circle:



Comments:

- In the images above, the dots are equally spaced, this need not indicate the data values are actually equally spaced as we are only interested in listing them in order.
- In fact, in the above pictures, two subsequent dots could have exactly the same value.
- It is clear that the value of the median will be in the same position regardless of the distance between data values.

Did I Get This?: Median

EXAMPLE: Median(2)

To find the median age of the Best Actress Oscar winners, we first need to order the data.

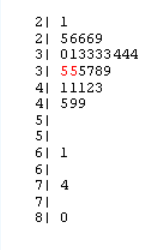
It would be useful, then, to use the stemplot, a diagram in which the data are already ordered.

- Here n = 32 (an even number), so the median M, will be the mean of the two center observations
- These are located at the (n / 2) = 32 / 2 = **16th** and (n / 2) + 1 = (32 / 2) + 1 = **17th**

Counting from the top, we find that:

- the 16th ranked observation is 35
- the 17th ranked observation also happens to be 35

Therefore, the median M = (35 + 35) / 2 = 35



Learn By Doing: Measures of Center #1

Comparing the Mean and the Median

Learning Objectives

LO 4.10: Choose the appropriate measures for a quantitative variable based upon the shape of the distribution.

Note

As we have seen, the **mean** and the **median**, the most common **measures of center**, each describe the center of a distribution of values in a different way.

- The mean describes the center as an average value, in which the **actual values**of the data points play an important role.
- The median, on the other hand, locates the middle value as the center, and the **order**of the data is the key.

To get a deeper understanding of the differences between these two measures of center, consider the following example. Here are two datasets:

Data set A → 64 65 66 68 70 71 **73**
Data set B → 64 65 66 68 70 71 **730**

For dataset A, the mean is 68.1, and the median is 68.

Looking at dataset B, notice that all of the observations except the last one are close together. The observation 730 is very large, and is certainly an outlier.

In this case, the median is still 68, but the mean will be influenced by the high outlier, and shifted up to 162.

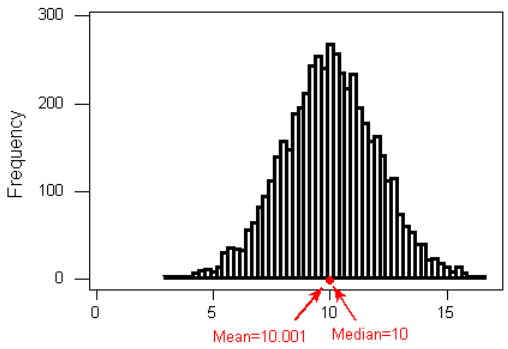
The message that we should take from this example is:

The mean is very sensitive to outliers (because it factors in their magnitude), while the median is resistant (or robust) to outliers.

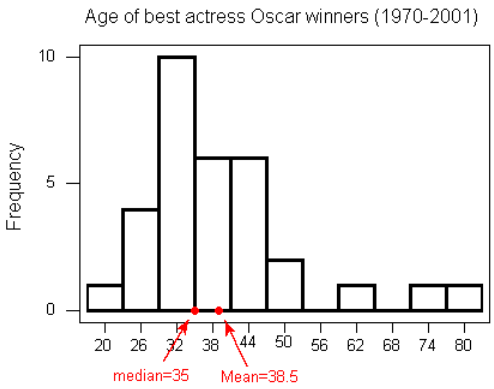
Interactive Applet: Comparing the Mean and Median

Therefore:

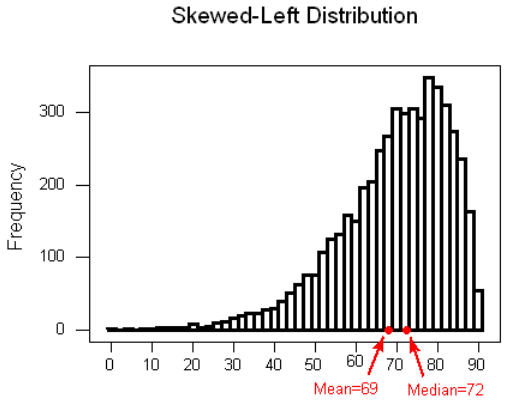
- For symmetric distributions with no outliers: the mean is approximately equal to the median.



- For skewed right distributions and/or datasets with high outliers: the mean is greater than the median.



- For skewed left distributions and/or datasets with low outliers: the mean is less than the median.



Conclusions… When to use which measures?

- Use the sample mean as a measure of center for symmetric distributions with no outliers.
- Otherwise, the median will be a more appropriate measure of the center of our data.

Did I Get This?: Measures of Center

Learn By Doing: Measures of Center #2

Learn By Doing: Measures of Center – Additional Practice

Let's Summarize

- The two main numerical measures for the center of a distribution are the mean and the median. The mean is the average value, while the median is the middle value.
- The mean is very sensitive to outliers (as it factors in their magnitude), while the median is resistant to outliers.
- The mean is an appropriate measure of center for symmetric distributions with no outliers. In all other cases, the median is often a better measure of the center of the distribution.

Measures of Spread

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Learning Objectives

LO 4.7: Define and describe the features of the distribution of one quantitative variable (shape, center, spread, outliers).

Video

Video: Measures of Spread (3 videos, 8:44 total)

Note

Related SAS Tutorials

- 5A – (3:01) Numeric Measures using PROC MEANS

Related SPSS Tutorials

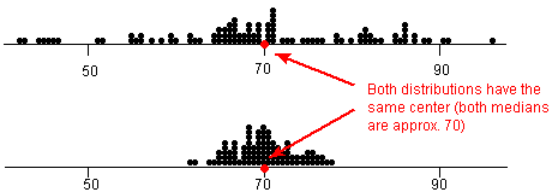
- 5A – (8:00) Numeric Measures using EXPLORE

Introduction

So far we have learned about different ways to quantify the center of a distribution. A measure of center by itself is not enough, though, to describe a distribution.

Consider the following two distributions of exam scores. Both distributions are centered at 70 (the median of both distributions is approximately 70), but the distributions are quite different.

The first distribution has a much larger variability in scores compared to the second one.



In order to describe the distribution, we therefore need to supplement the graphical display not only with a measure of center, but also with a measure of the variability (or spread) of the distribution.

In this section, we will discuss the three most commonly used measures of spread:

- Range
- Inter-quartile range (IQR)
- Standard deviation

Although the **measures of center** did approach the question differently, they do **attempt to measure the same point in the distribution** and thus are comparable.

However, the three **measures of spread** provide very different ways to quantify the variability of the distribution and **do not try to estimate the same quantity**.

In fact, the three **measures of spread provide information about three different aspects of the spread** of the distribution which, together, give a more complete picture of the spread of the distribution.

Range

Learning Objectives

LO 4.11: Define and calculate the range of one quantitative variable.

The **range** covered by the data is the most intuitive measure of variability. The range is exactly the distance between the smallest data point (min) and the largest one (Max).

- Range = Max – min

Note: When we first looked at the histogram, and tried to get a first feel for the spread of the data, we were actually approximating the range, rather than calculating the exact range.

EXAMPLE: Best Actress Oscar Winners

Here we have the Best Actress Oscar winners' data

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33

In this example:

- min = 21 (Marlee Matlin for *Children of a Lesser God*, 1986)
- Max = 80 (Jessica Tandy for *Driving Miss Daisy*, 1989)

The range covered by all the data is 80 – 21 = 59 years.

Inter-Quartile Range (IQR)

Learning Objectives

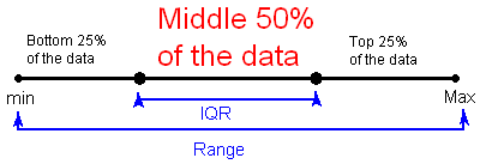
LO 4.12: Define and calculate Q1, Q3, and the IQR for one quantitative variable

While the range quantifies the variability by looking at the range covered by ALL the data,
the **Inter-Quartile Range** or **IQR** measures the variability of a distribution by giving us the range covered by the MIDDLE 50% of the data.

- **IQR** = Q3 – Q1
- **Q3** = 3rd Quartile = 75th Percentile
- **Q1** = 1st Quartile = 25th Percentile

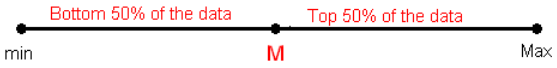
The following picture illustrates this idea: (Think about the horizontal line as the data ranging from the min to the Max). **IMPORTANT NOTE: The "lines" in the following illustrations are not to scale. The equal distances indicate equal amounts of data NOT equal distance between the numeric values.**

Although we will use software to calculate the quartiles and IQR, we will illustrate the basic process to help you fully understand.

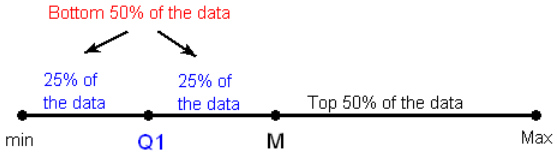


To calculate the IQR:

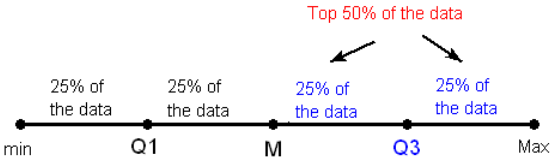
1. Arrange the data in increasing order, and find the median M. Recall that the median divides the data, so that 50% of the data points are below the median, and 50% of the data points are above the median.



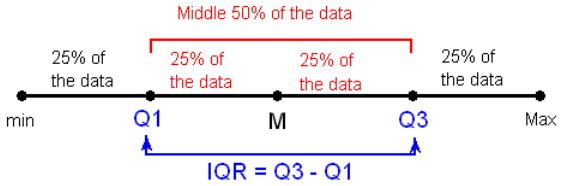
2. Find the median of the lower 50% of the data. This is called the first quartile of the distribution, and the point is denoted by Q1. Note from the picture that Q1 divides the lower 50% of the data into two halves, containing 25% of the data points in each half. Q1 is called the first quartile, since one quarter of the data points fall below it.



3. Repeat this again for the top 50% of the data. Find the median of the top 50% of the data. This point is called the third quartile of the distribution, and is denoted by Q3.
Note from the picture that Q3 divides the top 50% of the data into two halves, with 25% of the data points in each.Q3 is called the third quartile, since three quarters of the data points fall below it.



4. The middle 50% of the data falls between Q1 and Q3, and therefore: IQR = Q3 – Q1

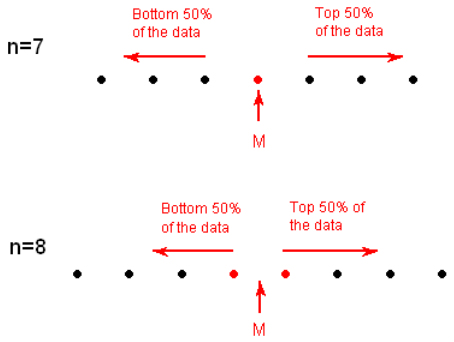


Comments:

1. The last picture shows that Q1, M, and Q3 divide the data into four quarters with 25% of the data points in each, where the median is essentially the second quartile. The use of IQR = Q3 – Q1 as a measure of spread is therefore particularly appropriate when the median M is used as a measure of center.

---

2. We can define a bit more precisely what is considered the bottom or top 50% of the data. The bottom (top) 50% of the data is all the observations whose position in the ordered list is to the left (right) of the location of the overall median M. The following picture will visually illustrate this for the simple cases of n = 7 and n = 8.



Note that when n is **odd** (as in n = 7 above), the median is **not** included in either the bottom or top half of the data; When n is **even** (as in n = 8 above), the data are naturally divided into two halves.

EXAMPLE: Best Actress Oscar Winners

To find the IQR of the Best Actress Oscar winners' distribution, it will be convenient to use the stemplot.

```
2| 1
2| 56669        Bottom Half
3| 013333444
3| 555789
4| 11123
4| 599
5|
5|
6| 1            Top half
6|
7| 4
7|
8| 0
```

Q1 is the median of the bottom half of the data. Since there are 16 observations in that half, Q1 is the mean of the 8th and 9th ranked observations in that half:

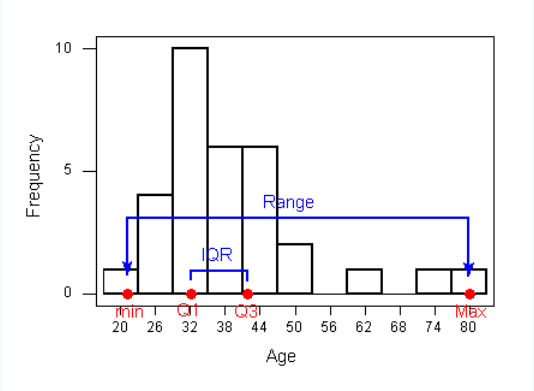
Q1 = (31 + 33) / 2 = 32

Similarly, Q3 is the median of the top half of the data, and since there are 16 observations in that half, Q3 is the mean of the 8th and 9th ranked observations in that half:

Q3 = (41 + 42) / 2 = 41.5

IQR = 41.5 – 32 = 9.5

Note that in this example, the range covered by all the ages is 59 years, while the range covered by the middle 50% of the ages is only 9.5 years. While the whole dataset is spread over a range of 59 years, the middle 50% of the data is packed into only 9.5 years. Looking again at the histogram will illustrate this:



Comment:

- Software packages use different formulas to calculate the quartiles Q1 and Q3. This should not worry you, as long as you understand the idea behind these concepts. For example, here are the quartile values provided by three different software packages for the age of best actress Oscar winners:

R:

```
> summary(actress)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   32.50   35.00   38.53   41.25   80.00
```

Minitab:

Descriptive Statistics: Age

Variable	N	Mean	Median	TrMean	StDev	SE Mean
actress	32	38.53	35.00	36.89	12.95	2.29

Variable	Minimum	Maximum	Q1	Q3
actress	21.00	80.00	**31.50**	**41.75**

Excel:

Q1	32.5
Q3	41.25

Q1 and Q3 as reported by the various software packages differ from each other and are also slightly different from the ones we found here. This should not worry you.

There are different acceptable ways to find the median and the quartiles. These can give different results occasionally, especially for datasets where n (the number of observations) is fairly small.

As long as you know what the numbers mean, and how to interpret them in context, it doesn't really matter much what method you use to find them, since the differences are negligible.

Standard Deviation

Learning Objectives

LO 4.13: Define and calculate the standard deviation and variance of one quantitative variable.

So far, we have introduced two measures of spread; the range (covered by all the data) and the inter-quartile range (IQR), which looks at the range covered by the middle 50% of the distribution. We also noted that the IQR should be paired as a measure of spread with the median as a measure of center.

We now move on to another measure of spread, the **standard deviation**, which quantifies the spread of a distribution in a completely different way.

Idea

The idea behind the standard deviation is to quantify the spread of a distribution by measuring how far the observations are from their mean. The standard deviation gives the average (or typical distance) between a data point and the mean.

Notation

There are many notations for the standard deviation: SD, s, Sd, StDev. Here, we'll use **SD** as an abbreviation for standard deviation, and use s as the symbol.

Formula

The **sample standard deviation formula** is:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

where,

s = sample standard deviation
n = number of scores in sample
$\sum$ = sum of...
and
$\overline{\mathcal{A}}$ = sample mean

Calculation

In order to get a better understanding of the standard deviation, it would be useful to see an example of how it is calculated. In practice, we will use a computer to do the calculation.

EXAMPLE: Video Store Customers

The following are the number of customers who entered a video store in 8 consecutive hours:

7, 9, 5, 13, 3, 11, 15, 9

To find the standard deviation of the number of hourly customers:

1. **Find the mean, x-bar, of your data:**

(7 + 9 + 5 + 13 + 3 + 11 + 15 + 9)/8 = 9

2. **Find the deviations from the mean:**

- The differences between each observation and the mean here are

(7 – 9), (9 – 9), (5 – 9), (13 – 9), (3 – 9), (11 – 9), (15 – 9), (9 – 9)

-2, 0, -4, 4, -6, 2, 6, 0

- Since the standard deviation attempts to measure the average (typical) distance between the data points and their mean, it would make sense to average the deviation we obtained.
- **Note,** however**, that the sum of the deviations is zero.**
- This is always the case, and is the reason why we need a more complex calculation.

3. **To solve the previous problem, in our calculation, we square each of the deviations.**

$(-2)^2$, $(0)^2$, $(-4)^2$, $(4)^2$, $(-6)^2$, $(2)^2$, $(6)^2$, $(0)^2$

4, 0, 16, 16, 36, 4, 36, 0

4. **Sum the squared deviations and divide by *n* – 1:**

$$(4 + 0 + 16 + 16 + 36 + 4 + 36 + 0)/(8 – 1)$$

$$(112)/(7) = 16$$

- The reason we divide by n-1 will be discussed later.
- This value, the sum of the squared deviations divided by n – 1, is called the **variance**. However, the variance is not used as a measure of spread directly as the units are the square of the units of the original data.

5. **The standard deviationof the data is the square root of the variance calculated in step 4:**

- In this case, we have the square root of 16 which is 4. We will use the lower case letter *s*to represent the standard deviation.

$$\boldsymbol{s = 4}$$

- We take the square root to obtain a measure which is in the original units of the data. The units of the variance of 16 are in "squared customers" which is difficult to interpret.
- The units of the standard deviation are in "customers" which makes this measure of variation more useful in practice than the variance.

Recall that the average of the number of customers who enter the store in an hour is 9.

The interpretation of the standard deviation is that on average, the actual number of customers who enter the store each hour is 4 away from 9.

Comment: The importance of the numerical figure that we found in #4 above called the variance (=16 in our example) will be discussed much later in the course when we get to the inference part.

Learn By Doing: Standard Deviation

Properties of the Standard Deviation

1. It should be clear from the discussion thus far that the SD should be paired as a measure of spread with the mean as a measure of center.

---

2. Note that the only way, mathematically, in which the SD = 0, is when all the observations have the same value (Ex: 5, 5, 5, … , 5), in which case, the deviations from the mean (which is also 5) are all 0. This is intuitive, since if all the data points have the same value, we have no variability (spread) in the data, and expect the measure of spread (like the SD) to be 0. Indeed, in this case, not only is the SD equal to 0, but the range and the IQR are also equal to 0. Do you understand why?

---

3. Like the mean, the SD is strongly influenced by outliers in the data. Consider the example concerning video store customers: 3, 5, 7, 9, 9, 11, 13, 15 (data ordered). If the largest observation was wrongly recorded as 150, then the average would jump up to 25.9, and the standard deviation would jump up to SD = 50.3. Note that in this simple example, it is easy to see that while the standard deviation is strongly influenced by outliers, the IQR is not. The IQR would be the same in both cases, since, like the median, the calculation of the quartiles depends only on the order of the data rather than the actual values.

The last comment leads to the following very important conclusion:

Choosing Numerical Measures

Learning Objectives

LO 4.10: Choose the appropriate measures for a quantitative variable based upon the shape of the distribution.

- Use the **mean and the standard deviation** as measures of center and spread for **reasonably symmetric distributions with no extreme outliers.**
- **For all other cases**, use **the five-number summary = min, Q1, Median, Q3, Max** (which gives the median, and easy access to the IQR and range). We will discuss the five-number summary in the next section in more detail.

Let's Summarize

- The **range** covered by the data is the most intuitive measure of spread and is exactly the distance between the smallest data point (min) and the largest one (Max).
- Another measure of spread is the **inter-quartile range (IQR)**, which is the range covered by the middle 50% of the data.
- IQR = Q3 – Q1, the difference between the third and first quartiles.
  - The **first quartile (Q1)** is the value such that one quarter (25%) of the data points fall below it, or the median of the bottom half of the data.
  - The **third quartile (Q3)** is the value such that three quarters (75%) of the data points fall below it, or the median of the top half of the data.
- The **IQR** is generally used as a measure of spread of a distribution when the **median** is used as a measure of center.
- The **standard deviation** measures the spread by reporting **a typical (average) distance between the data points and their mean.**
- It is appropriate to use the **standard deviation** as a measure of spread with the **mean** as the measure of center.
- Since the **mean and standard deviations are highly influenced by extreme observations**, they should be used as numerical descriptions of the center and spread **only for distributions that are roughly symmetric, and have no extreme outliers. In all other situations, we prefer the 5-number summary.**

Measures of Position

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Learning Objectives

LO 4.14: Define and interpret measures of position (percentiles, quartiles, the five-number summary, z-scores).

Video

Video: Measures of Position (2 videos, 4:20 total)

Note

Related SAS Tutorials

- 5A – (3:01) Numeric Measures using PROC MEANS

Related SPSS Tutorials

- 5A – (8:00) Numeric Measures using EXPLORE

Although not a required aspect of describing distributions of one quantitative variable, we are often interested in where a particular value falls in the distribution. Is the value unusually low or high or about what we would expect?

Answers to these questions rely on measures of position (or location). These measures give information about the distribution but also give information about how individual values relate to the overall distribution.

Percentiles

A common measure of position is the percentile. Although there are some mathematical considerations involved with calculating percentiles which we will not discuss, you should have a basic understanding of their interpretation.

> In general the ***P*-th percentile** can be interpreted as a location in the data for which approximately *P*% of the other values in the distribution fall below the *P*-th percentile and (100 –*P*)% fall above the *P*-th percentile.

The quartiles Q1 and Q3 are special cases of percentiles and thus are measures of position.

Five-Number Summary

The combination of the five numbers (min, Q1, M, Q3, Max) is called the **five number summary**, and provides a quick numerical description of both the center and spread of a distribution.

Each of the values represents a measure of position in the dataset.

The min and max providing the boundaries and the quartiles and median providing information about the 25th, 50th, and 75th percentiles.

Standardized Scores (Z-Scores)

Standardized scores, also called z-scores use the mean and standard deviation as the primary measures of center and spread and are therefore most useful when the mean and standard deviation are appropriate, i.e. when the distribution is reasonably symmetric with no extreme outliers.

For any individual, the **z-score** tells us how many standard deviations the raw score for that individual deviates from the mean and in what direction. A positive z-score indicates the individual is above average and a negative z-score indicates the individual is below average.

To calculate a z-score, we take the individual value and subtract the mean and then divide this difference by the standard deviation.

$$z_i = \frac{x_i - \bar{x}}{S}$$

Measures of Position

Measures of position also allow us to compare values from different distributions. For example, we can present the percentiles or z-scores of an individual's height and weight. These two measures together would provide a better picture of how the individual fits in the overall population than either would alone.

Although measures of position are not stressed in this course as much as measures of center and spread, we have seen and will see many measures of position used in various aspects of examining the distribution of one variable and it is good to recognize them as measures of position when they appear.

Outliers

> **CO-4:** Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Learning Objectives

LO 4.7: Define and describe the features of the distribution of one quantitative variable (shape, center, spread, outliers).

Video

Video: Outliers (2:30)

Using the IQR to Detect Outliers

Learning Objectives

LO 4.15: Define and use the 1.5(IQR) and 3(IQR) criterion to identify potential outliers and extreme outliers.

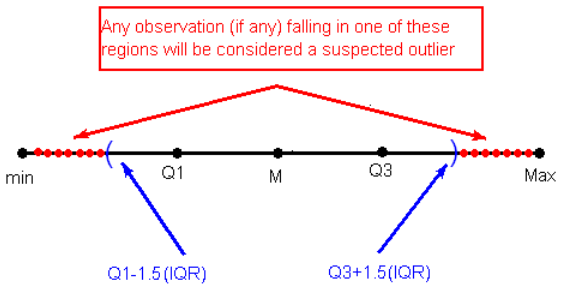
So far we have quantified the idea of center, and we are in the middle of the discussion about measuring spread, but we haven't really talked about a method or rule that will help us classify extreme observations as outliers. The IQR is commonly used as the basis for a rule of thumb for identifying outliers.

The 1.5(IQR) Criterion for Outliers

An observation is considered a **suspected outlier** or **potential outlier** if it is:

- below Q1 – 1.5(IQR) or
- above Q3 + 1.5(IQR)

The following picture (not to scale) illustrates this rule:



EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example (Link to the Best Actress Oscar Winners data).

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33

Recall that when we first looked at the histogram of ages of Best Actress Oscar winners, there were three observations that looked like possible outliers:

We can now use the 1.5(IQR) criterion to check whether the three highest ages should indeed be classified as potential outliers:

- For this example, we found Q1 = 32 and Q3 = 41.5 which give an IQR = 9.5
- Q1 – 1.5 (IQR) = 32 – (1.5)(9.5) = 17.75

- Q3 + 1.5 (IQR) = 41.5 + (1.5)(9.5) = 55.75

The 1.5(IQR) criterion tells us that any observation with an age that is below 17.75 or above 55.75 is considered a suspected outlier.

We therefore conclude that the observations with ages of 61, 74 and 80 should be flagged as suspected outliers in the distribution of ages. Note that since the smallest observation is 21, there are no suspected low outliers in this distribution.

The 3(IQR) Criterion for Outliers

An observation is considered an **EXTREME outlier** if it is:

- below Q1 – 3(IQR) or
- above Q3 + 3(IQR)

EXAMPLE: Best Actress Oscar Winners

We can now use the 3(IQR) criterion to check whether any of the three suspected outliers can be classified as extreme outliers:

- For this example, we found Q1 = 32 and Q3 = 41.5 which give an IQR = 9.5
- Q1 – 3 (IQR) = 32 – (3)(9.5) = 3.5
- Q3 + 3 (IQR) = 41.5 + (3)(9.5) = 70

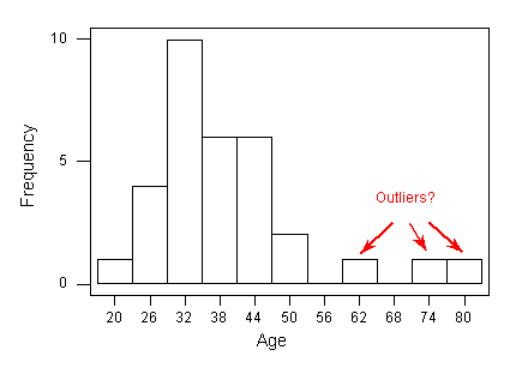
The 3(IQR) criterion tells us that any observation that is below 3.5 or above 70 is considered an extreme outlier.

We therefore conclude that the observations with ages 74 and 80 should be flagged as extreme outliers in the distribution of ages.

Note that since there were no suspected outliers on the low end there can be no extreme outliers on the low end of the distribution. Thus there was no real need for us to calculate the low cutoff for extreme outliers, i.e. Q1 – 3(IQR) = 3.5.

See the histogram below, and consider the outliers individually.

- The observation with age 62 is visually much closer to the center of the data. We might have a difficult time deciding if this value is really an outlier using this graph alone.
- However, the ages of 74 and 80 are clearly far from the bulk of the distribution. We might feel very comfortable deciding these values are outliers based only on the graph.



Did I Get This?: Identifying Outliers using IQR Method

Understanding Outliers

Learning Objectives

LO 4.16: Discuss possible methods for handling outliers in practice.

We just practiced one way to 'flag' possible outliers. Why is it important to identify possible outliers, and how should they be dealt with? The answers to these questions depend on the reasons for the outlying values. Here are several possibilities:

1. Even though it is an extreme value, if an outlier can be understood to have been produced by **essentially the same sort of physical or biological process** as the rest of the data, and if such extreme values are expected to **eventually occur again**, then such an outlier indicates something important and interesting about the process you're investigating, and it **should be kept** in the data.

---

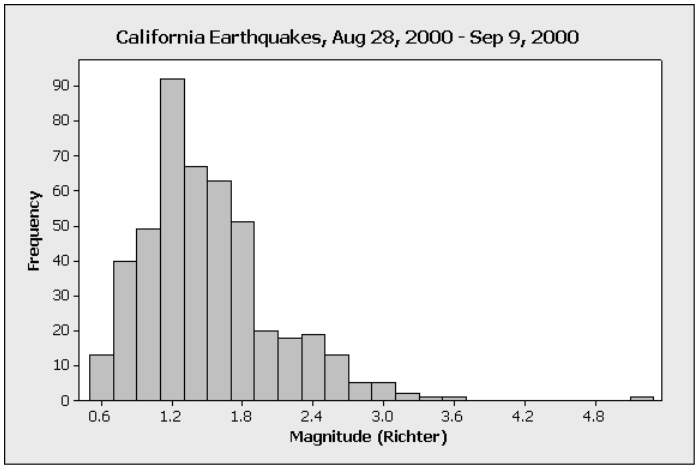
2. If an outlier can be explained to have been produced under fundamentally **different** conditions from the rest of the data (or by a fundamentally different process), such an outlier **can be removed** from the data if your goal is to investigate only the process that produced the rest of the data.

---

3. An outlier might indicate a **mistake** in the data (like a typo, or a measuring error), in which case it **should be corrected if possible or else removed** from the data before calculating summary statistics or making inferences from the data (and the reason for the mistake should be investigated).

Here are examples of each of these types of outliers:

1. The following histogram displays the magnitude of 460 earthquakes in California, occurring in the year 2000, between August 28 and September 9:



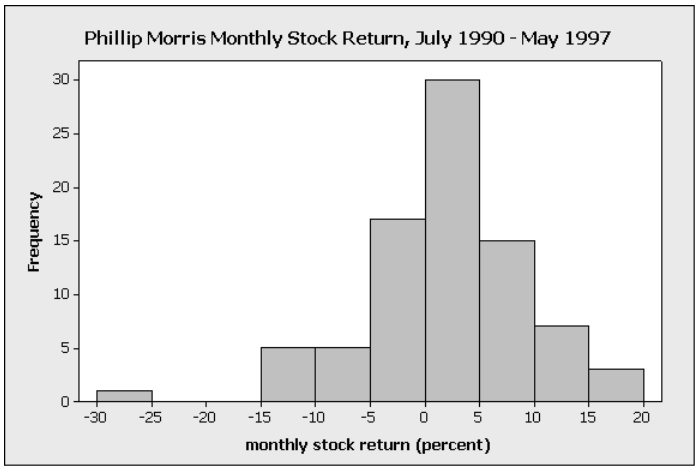
Identifying the outlier: On the very far right edge of the display (beyond 4.8), we see a low bar; this represents one earthquake (because the bar has height of 1) that was much more severe than the others in the data.

Understanding the outlier: In this case, the outlier represents a much stronger earthquake, which is relatively rarer than the smaller quakes that happen more frequently in California.

How to handle the outlier: For many purposes, the relatively severe quakes represented by the outlier might be the most important (because, for instance, that sort of quake has the potential to do more damage to people and infrastructure). The smaller-magnitude quakes might not do any damage, or even be felt at all. So, for many purposes it could be important to keep this outlier in the data.

---

2. The following histogram displays the monthly percent return on the stock of Phillip Morris (a large tobacco company) from July 1990 to May 1997:



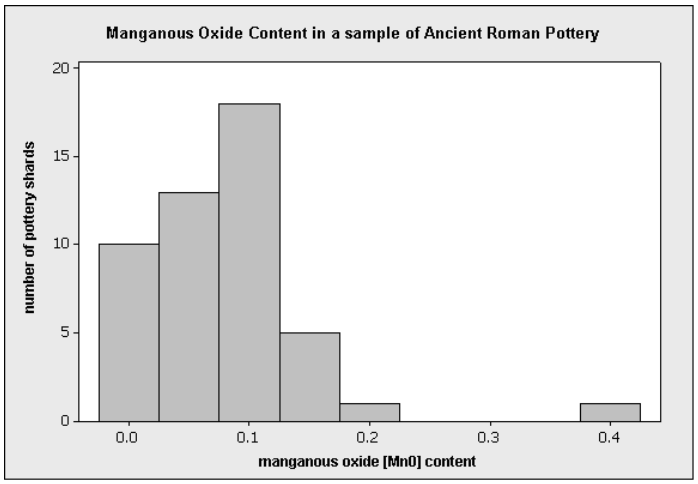
Identifying the outlier: On the display, we see a low bar far to the left of the others; this represents one month's return (because the bar has height of 1), where the value of Phillip Morris stock was unusually low.

Understanding the outlier: The explanation for this particular outlier is that, in the early 1990s, there were highly-publicized federal hearings being conducted regarding the addictiveness of smoking, and there was growing public sentiment against the tobacco companies. The unusually low monthly value in the Phillip Morris dataset was due to public pressure against smoking, which negatively affected the company's stock for that particular month.

How to handle the outlier: In this case, the outlier was due to unusual conditions during one particular month that aren't expected to be repeated, and that were fundamentally different from the conditions that produced the values in all the other months. So in this case, it would be reasonable to remove the outlier, if we wanted to characterize the "typical" monthly return on Phillip Morris stock.

---

3. When archaeologists dig up objects such as pieces of ancient pottery, chemical analysis can be performed on the artifacts. The chemical content of pottery can vary depending on the type of clay as well as the particular manufacturing technique. The following histogram displays the results of one such actual chemical analysis, performed on 48 ancient Roman pottery artifacts from archaeological sites in Britain:

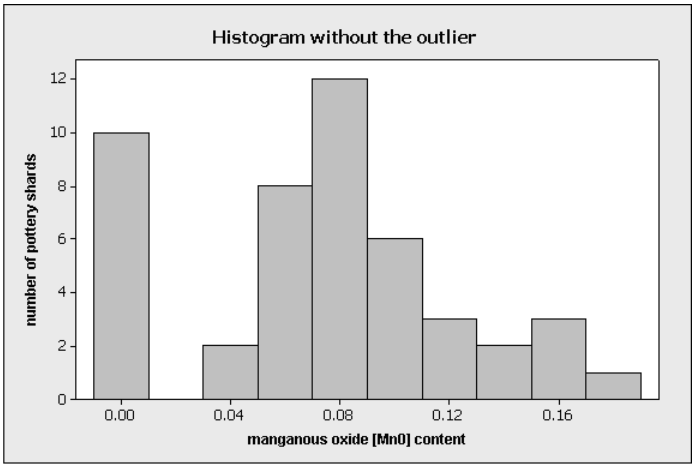


As appeared in Tubb, et al. (1980). "The analysis of Romano-British pottery by atomic absorption spectrophotometry." Archaeometry, vol. 22, reprinted in Statistics in Archaeology by Michael Baxter, p. 21.

Identifying the outlier: On the display, we see a low bar far to the right of the others; this represents one piece of pottery (because the bar has a height of 1), which has a suspiciously high manganous oxide value.

Understanding the outlier: Based on comparison with other pieces of pottery found at the same site, and based on expert understanding of the typical content of this particular compound, it was concluded that the unusually high value was most likely a typo that was made when the data were published in the original 1980 paper (it was typed as ".394" but it was probably meant to be ".094").

How to handle the outlier: In this case, since the outlier was judged to be a mistake, it should be removed from the data before further analysis. In fact, removing the outlier is useful not only because it's a mistake, but also because doing so reveals important structure that was otherwise hidden. This feature is evident on the next display:



When the outlier is removed, the display is re-scaled so that now we can see the set of 10 pottery pieces that had almost no manganous oxide. These 10 pieces might have been made with a different potting technique, so identifying them as different from the rest is historically useful. This feature was only evident after the outlier was removed.

Reading: Outliers (≈ 1400 words)

Boxplots

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Learning Objectives

LO 4.7: Define and describe the features of the distribution of one quantitative variable (shape, center, spread, outliers).

Video

Video: Boxplots (2 videos, 7:02 total)

Note

Related SAS Tutorials

- 5B – (4:05) Creating Histograms and Boxplots using SGPLOT

Related SPSS Tutorials

- 5B – (2:29) Creating Histograms and Boxplots

Introduction

Now we introduce another graphical display of the distribution of a quantitative variable, the **boxplot**.

The Five Number Summary

So far, in our discussion about measures of spread, some key players were:

- the extremes (min and Max), which provide the range covered by all the data; and
- the quartiles (Q1, M and Q3), which together provide the IQR, the range covered by the middle 50% of the data.

Recall that the combination of all five numbers (min, Q1, M, Q3, Max) is called the **five number summary**, and provides a quick numerical description of both the center and spread of a distribution.

EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example (Link to the Best Actress Oscar Winners data).

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33

The five number summary of the age of Best Actress Oscar winners (1970-2001) is:

min = 21, Q1 = 32, M = 35, Q3 = 41.5, Max = 80

To sketch the boxplot we will need to know the 5-number summary as well as identify any outliers. We will also need to locate the largest and smallest values which are not outliers. The stemplot below might be helpful as it displays the data in order.

```
   4
   4
   4
   3 9
  9 3 8 3
  6 3 7 2
  6 3 5 1 9
  6 1 5 1 9
1 5 0 5 1 5     1   4   0
-------------------------
2 2 3 3 4 4 5 5 6 6 7 7 8
```

Learn By Doing: 5-Number Summary

Now that you understand what each of the five numbers means, you can appreciate how much information about the distribution is packed into the five-number summary. All this information can also be represented visually by using the boxplot.

The Boxplot

Learning Objectives

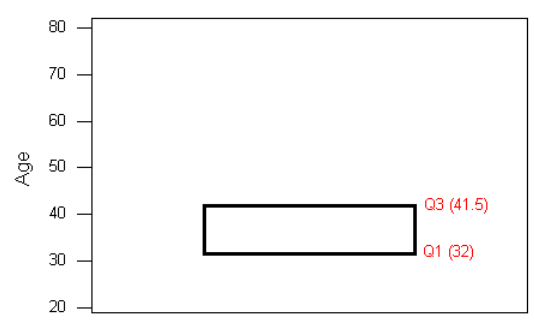
LO 4.17: Explain the process of creating a boxplot (including appropriate indication of outliers).

The boxplot graphically represents the distribution of a quantitative variable by visually displaying the five-number summary and any observation that was classified as a suspected outlier using the 1.5(IQR) criterion.

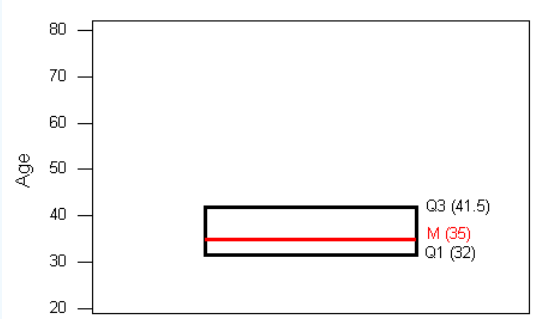
EXAMPLE: Constructing a boxplot

(Link to the Best Actress Oscar Winners data).

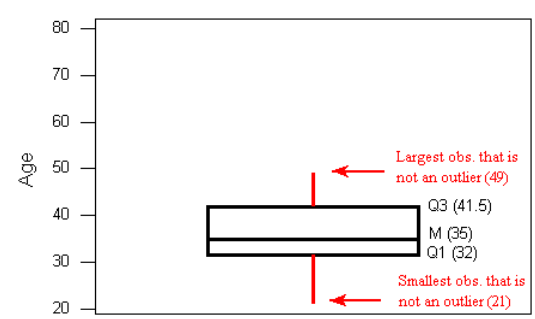
1. The central box spans from Q1 to Q3. In our example, the box spans from 32 to 41.5. Note that the width of the box has no meaning.



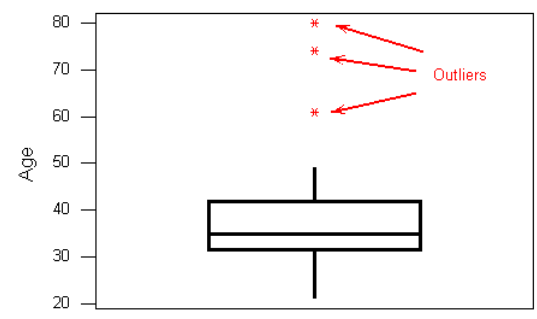
2. A line in the box marks the median M, which in our case is 35.



3. Lines extend from the edges of the box to the smallest and largest observations that were not classified as suspected outliers (using the 1.5xIQR criterion). In our example, we have no low outliers, so the bottom line goes down to the smallest observation, which is 21. Since we have three high outliers (61,74, and 80), the top line extends only up to 49, which is the largest observation that has not been flagged as an outlier.

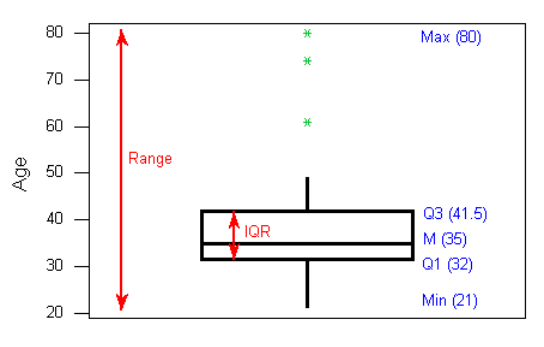


4. outliers are marked with asterisks (*).



To summarize: the following information is visually depicted in the boxplot:

- the five number summary (blue)
- the range and IQR (red)
- outliers (green)



Learn By Doing: Boxplots

Did I Get This?: Boxplots

Side-By-Side (Comparative) Boxplots

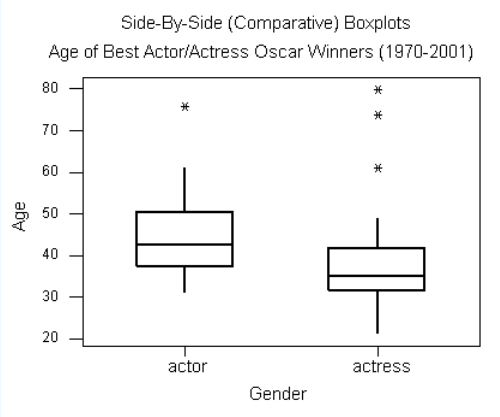
Learning Objectives

LO 4.18: Compare and contrast distributions (of quantitative data) from two or more groups, and produce a brief summary, interpreting your findings in context.

As we learned earlier, the distribution of a quantitative variable is best represented graphically by a histogram. Boxplots are most useful when presented side-by-side for comparing and contrasting distributions from two or more groups.

EXAMPLE: Best Actress/Actor Oscar Winners

So far we have examined the age distributions of Oscar winners for males and females separately. It will be interesting to compare the age distributions of actors and actresses who won best acting Oscars. To do that we will look at side-by-side boxplots of the age distributions by gender.



Recall also that we found the five-number summary and means for both distributions. For the Best Actress dataset, we did the calculations by hand. For the Best Actor dataset, we used statistical software, and here are the results:

- Actors: min = 31, Q1 = 37.25, M = 42.5, Q3 = 50.25, Max = 76
- Actresses: min = 21, Q1 = 32, M = 35, Q3 = 41.5, Max = 80

Based on the graph and numerical measures, we can make the following comparison between the two distributions:

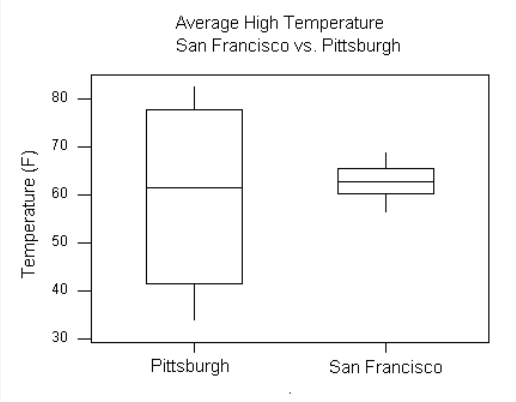
Center: The graph reveals that the age distribution of the males is higher than the females' age distribution. This is supported by the numerical measures. The median age for females (35) is lower than for males (42.5). Actually, it should be noted that even the third quartile of the females' distribution (41.5) is lower than the median age for males. We therefore conclude that in general, actresses win the Best Actress Oscar at a younger age than actors do.

Spread: Judging by the range of the data, there is much more variability in the females' distribution (range = 59) than there is in the males' distribution (range = 45). On the other hand, if we look at the IQR, which measures the variability only among the middle 50% of the distribution, we see more spread in the ages of males (IQR = 13) than females (IQR = 9.5). We conclude that among all the winners, the actors' ages are more alike than the actresses' ages. However, the middle 50% of the age distribution of actresses is more homogeneous than the actors' age distribution.

Outliers: We see that we have outliers in both distributions. There is only one high outlier in the actors' distribution (76, Henry Fonda, On Golden Pond), compared with three high outliers in the actresses' distribution.

EXAMPLE: Temperature of Pittsburg vs. San Francisco

In order to compare the average high temperatures of Pittsburgh to those in San Francisco we will look at the following side-by-side boxplots, and supplement the graph with the descriptive statistics of each of the two distributions.



Statistic	Pittsburgh	San Francisco
min	33.7	56.3
Q1	41.2	60.2
Median	61.4	62.7
Q3	77.75	65.35
Max	82.6	68.7

When looking at the graph, the similarities and differences between the two distributions are striking. Both distributions have roughly the same center (medians are 61.4 for Pitt, and 62.7 for San Francisco). However, the temperatures in Pittsburgh have a much larger variability than the temperatures in San Francisco (Range: 49 vs. 12. IQR: 36.5 vs. 5).

The practical interpretation of the results we obtained is that the weather in San Francisco is much more consistent than the weather in Pittsburgh, which varies a lot during the year. Also, because the temperatures in San Francisco vary so little during the year, knowing that the median temperature is around 63 is actually very informative. On the other hand, knowing that the median temperature in Pittsburgh is around 61 is practically useless, since temperatures vary so much during the year, and can get much warmer or much colder.

Note that this example provides more intuition about variability by interpreting small variability as consistency, and large variability as lack of consistency. Also, through this example we learned that the center of the distribution is more meaningful

as a typical value for the distribution when there is little variability (or, as statisticians say, little “noise”) around it. When there is large variability, the center loses its practical meaning as a typical value.

Learn By Doing: Comparing Distributions with Boxplots

Let’s Summarize

- The five-number summary of a distribution consists of the median (M), the two quartiles (Q1, Q3) and the extremes (min, Max).
- The five-number summary provides a complete numerical description of a distribution. The median describes the center, and the extremes (which give the range) and the quartiles (which give the IQR) describe the spread.
- The boxplot graphically represents the distribution of a quantitative variable by visually displaying the five number summary and any observation that was classified as a suspected outlier using the 1.5(IQR) criterion. (Some software packages indicate extreme outliers with a different symbol)
- Boxplots are most useful when presented side-by-side to compare and contrast distributions from two or more groups.

The "Normal" Shape

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 4.4: Using appropriate graphical displays and/or numerical measures, describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern

Learning Objectives

LO 4.7: Define and describe the features of the distribution of one quantitative variable (shape, center, spread, outliers).

Video

Video: The Normal Shape (5:34)

Related SAS Tutorials

- 5B – (4:05) Creating Histograms and Boxplots using SGPLOT
- 5C – (5:41) Creating QQ-Plots and other plots using UNIVARIATE

Related SPSS Tutorials

- 5B – (2:29) Creating Histograms and Boxplots
- 5C – (2:31) Creating QQ-Plots and PP-Plots

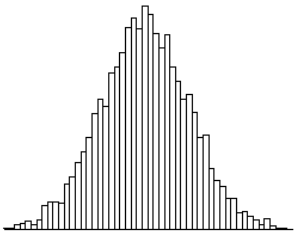
The Standard Deviation Rule

Learning Objectives

LO 6.2: Apply the standard deviation rule to the special case of distributions having the "normal" shape.

In the previous activity we tried to help you develop better intuition about the concept of standard deviation. The rule that we are about to present, called "The Standard Deviation Rule" (also known as "The Empirical Rule") will hopefully also contribute to building your intuition about this concept.

Consider a symmetric mound-shaped distribution:

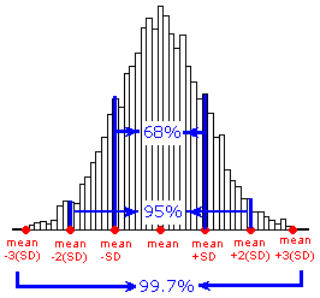


For distributions having this shape (later we will define this shape as "normally distributed"), the following rule applies:

The Standard Deviation Rule:

- Approximately 68% of the observations fall within 1 standard deviation of the mean.
- Approximately 95% of the observations fall within 2 standard deviations of the mean.
- Approximately 99.7% (or virtually all) of the observations fall within 3 standard deviations of the mean.

The following picture illustrates this rule:



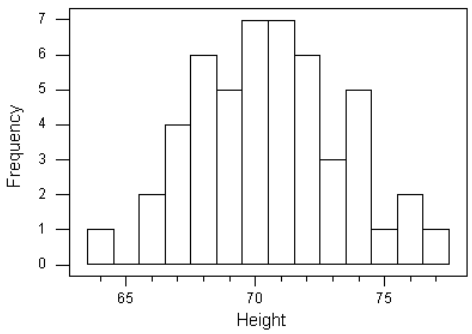
This rule provides another way to interpret the standard deviation of a distribution, and thus also provides a bit more intuition about it.

Interactive Applet: The Standard Deviation Rule

To see how this rule works in practice, consider the following example:

EXAMPLE: MALE HEIGHT

The following histogram represents height (in inches) of 50 males. Note that the data are roughly normal, so we would like to see how the Standard Deviation Rule works for this example.



Below are the actual data, and the numerical measures of the distribution. Note that the key players here, the mean and standard deviation, have been highlighted.

Statistic	Height
N	50
Mean	**70.58**
StDev	**2.858**
min	64
Q1	68
Median	70.5
Q3	72
Max	77

To see how well the Standard Deviation Rule works for this case, we will find what percentage of the observations falls within 1, 2, and 3 standard deviations from the mean, and compare it to what the Standard Deviation Rule tells us this percentage should be.

Interval	Mean-SD, Mean+SD (67.7 , 73.4)	Mean-2(SD),Mean+2(SD) (64.9 , 76.3)	Mean-3(SD), Mean+3(SD) (62 , 79.2)
Percentage of Observations in interval	34 observations 34/50 = 68%	48 observations 48/50 = 96%	All 50 observations 50/50 = 100%
SD Rule says...	68%	95%	99.7%

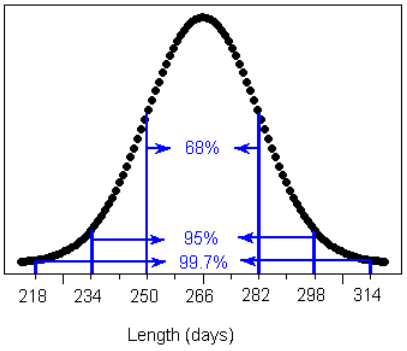
It turns out the Standard Deviation Rule works **very well** in this example.

The following example illustrates how we can apply the Standard Deviation Rule to variables whose distribution is known to be approximately normal.

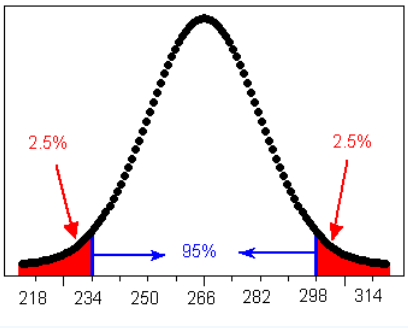
EXAMPLE: Length of Human Pregnancy

The length of the human pregnancy is not fixed. It is known that it varies according to a distribution which is roughly normal, with a mean of 266 days, and a standard deviation of 16 days. (Source: Figures are from Moore and McCabe, *Introduction to the Practice of Statistics*).

First, let's apply the Standard Deviation Rule to this case by drawing a picture:

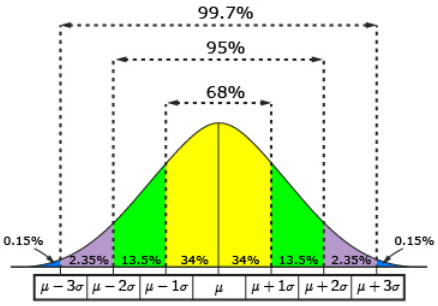


- Question: How long do the middle 95% of human pregnancies last? We can now use the information provided by the Standard Deviation Rule about the distribution of the length of human pregnancy, to answer some questions. For example:
  - Answer: The middle 95% of pregnancies last within 2 standard deviations of the mean, or in this case 234-298 days.
- Question: What percent of pregnancies last more than 298 days?
  - Answer: To answer this consider the following picture:



- Question: How short are the shortest 2.5% of pregnancies? Since 95% of the pregnancies last between 234 and 298 days, the remaining 5% of pregnancies last either less than 234 days or more than 298 days. Since the normal distribution is symmetric, these 5% of pregnancies are divided evenly between the two tails, and therefore 2.5% of pregnancies last more than 298 days.
  - Answer: Using the same reasoning as in the previous question, the shortest 2.5% of human pregnancies last less than 234 days.
- Question: What percent of human pregnancies last more than 266 days?
  - Answer: Since 266 days is the mean, approximately 50% of pregnancies last more than 266 days.

Here is a complete picture of the information provided by the standard deviation rule.



Did I Get This?: Standard Deviation Rule

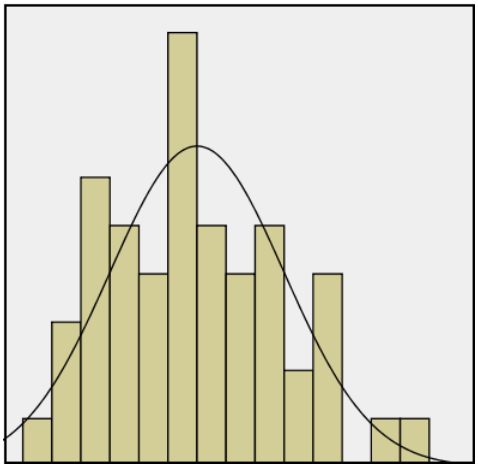
Visual Methods of Assessing Normality

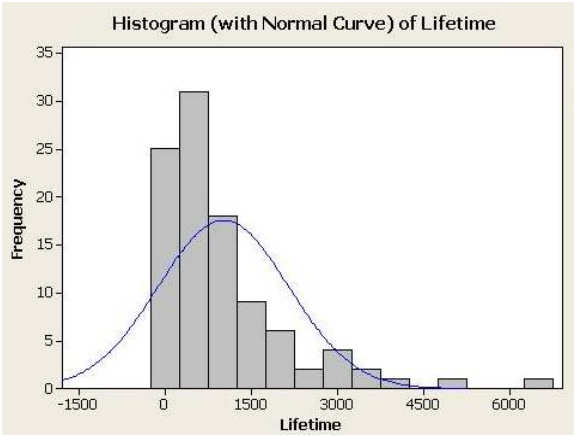
Learning Objectives

LO 6.3: Use histograms and QQ-plots (or Normal Probability Plots) to visually assess the normality of distributions of quantitative variables.

The normal distribution exists in theory but rarely, if ever, in real life. Histograms provide an excellent graphical display to help us assess normality. We can add a "normal curve" to the histogram which shows the normal distribution having the same mean and standard deviation as our sample. The closer the histogram fits this curve, the more (perfectly) normal the sample.

In the examples below, the graph on the top is approximately normally distributed whereas the graph on the bottom is clearly skewed right.

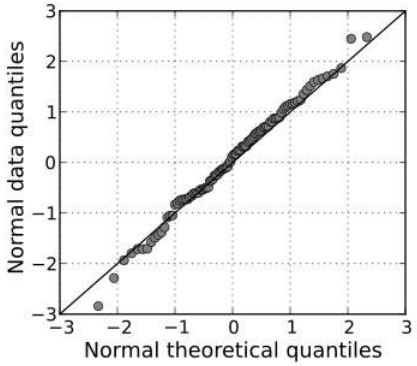




Unfortunately, we cannot quantitatively determine the extent to which the distribution is normally or not normally distributed using this method, but it can be helpful for making qualitative judgments about whether the data approximates the normal curve.

Another common graph to assess normality is the **Q-Q plot** (or **Normal Probability Plot**). In these graphs, the percentiles or quantiles of the theoretical distribution (in this case the standard normal distribution) are plotted against those from the data. If the data matches the theoretical distribution, the graph will result in a straight line. The graph below shows a distribution which closely follows a normal model.

Note: QQ-plots are not scatterplots (which we will dicuss soon), they only display information about one quantitative variable and graph this against the theoretical or expected values from a normal distribution with the same mean and standard deviation as our data. Other distributions can also be used.



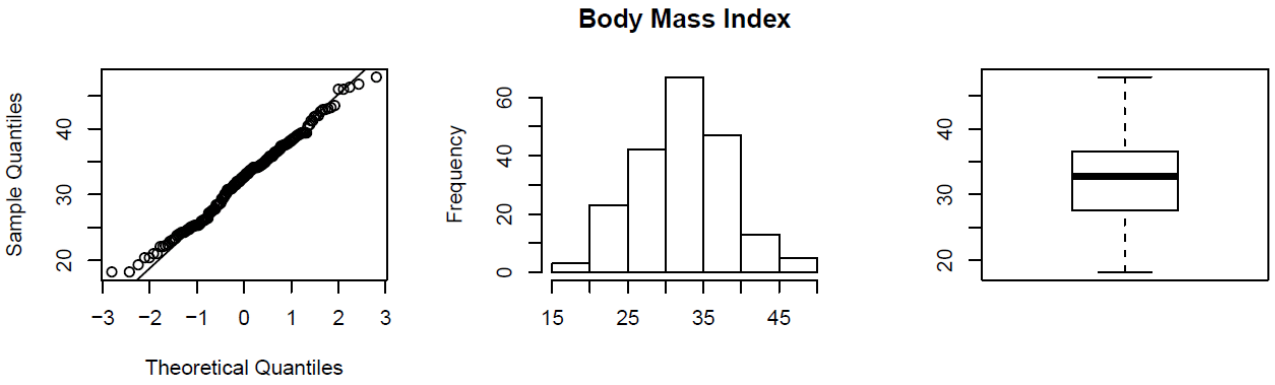
In most cases the distributions that you encounter will only be approximations of the normal curve, or they will not resemble the normal distribution at all! However, it can be important to consider how well the data being analyzed approximates the normal curve since this distribution is a key assumption of many statistical analyses.

Here are a few more examples:

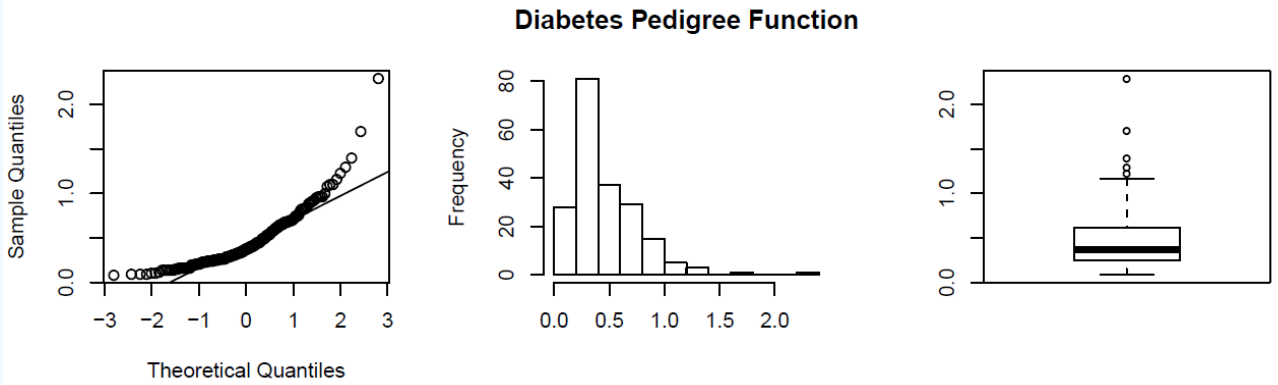
EXAMPLE: Some Real Data

The following gives the QQ-plot, histogram and boxplot for variables from a dataset from a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, who were tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. We used the 532 complete records after dropping the (mainly missing) data on serum insulin.

Body Mass Index is definitely **unimodal** and **symmetric** and could easily have come from a population which is **normally distributed**.

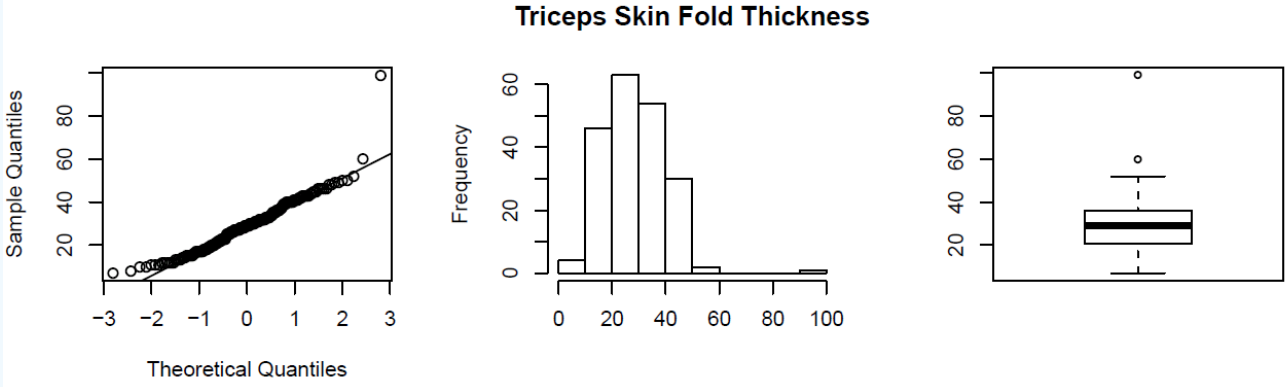


The Diabetes Pedigree Function scores were unimodal and skewed right. This data does not seem to have come from a population which is normally distributed.



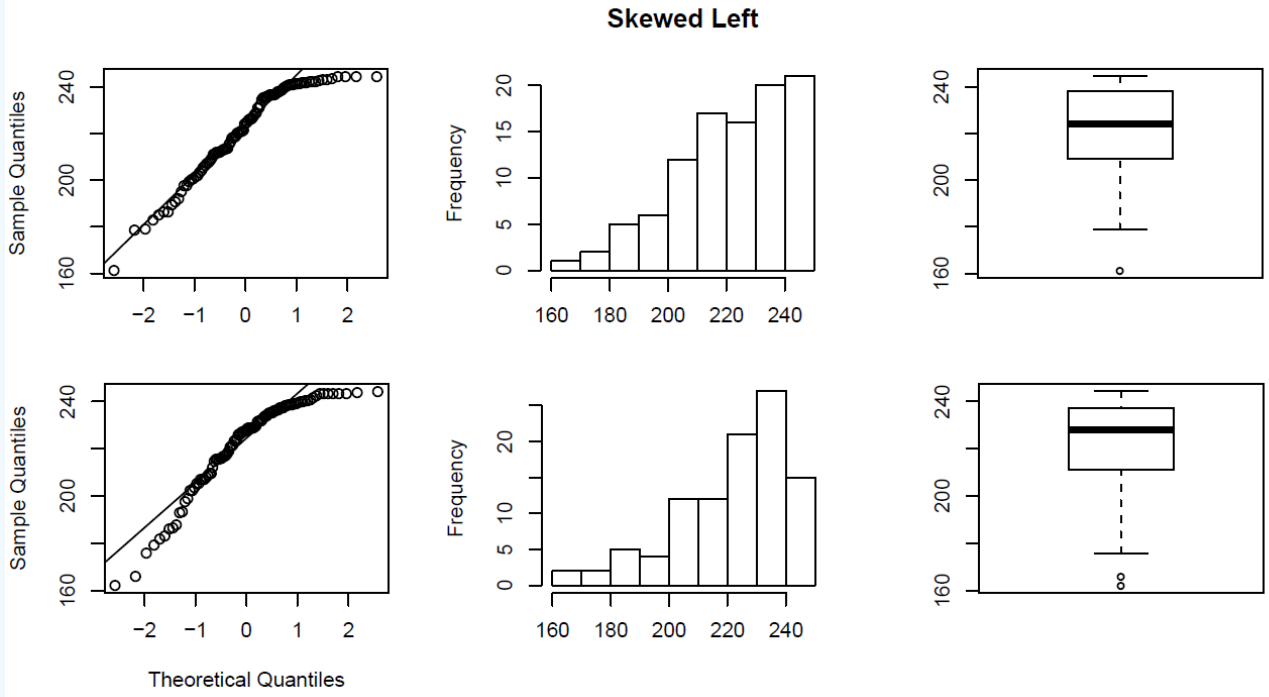
The Triceps Skin Fold Thickness is **basically symmetric with one extreme outlier** (and one potential but mild outlier).

Be careful not to call such a distribution "skewed right" as it is only the single outlier which really shows that pattern here. At a minimum remove the outlier and recreate the graphs to see how skewed the rest of the data might be.



EXAMPLE: Randomly Generated Data

Since there were no skewed left examples in the real data, here are two randomly generated skewed left distributions. Notice that the first is less skewed left than the second and this is indicated clearly in all three plots.



Comments:

- Even if the population is exactly normally distributed, samples from this population can appear non-normal especially for small sample sizes. See this document containing 21 samples of size n = 50 from a normal distribution with a mean of 200 and a standard deviation of 30. The samples that produce results which are skewed or otherwise seemingly not-normal are highlighted but even among those not highlighted, notice the variation in shapes seen: Normal Samples
- The standard deviation rule can also help in assessing normality in that the closer the percentage of data points within 1, 2, and 3 standard deviations is to that of the rule, the closer the data itself fits a normal distribution.
- In our example of male heights, we see that the histogram resembles a normal distribution and the sample percentages are very close to that predicted by the standard deviation rule.

Did I Get This?: Assessing Normality

(Optional) Reading: The Normal Distribution (≈ 500 words)

Standardized Scores (Z-Scores)

Learning Objectives

LO 4.14: Define and interpret measures of position (percentiles, quartiles, the five-number summary, z-scores).

We have already learned the standard deviation rule, which for normally distributed data, provides approximations for the proportion of data values within 1, 2, and 3 standard deviations. From this we know that approximately 5% of the data values would be expected to fall OUTSIDE 2 standard deviations.

If we calculate the standardized scores (or z-scores) for our data, it would be easy to identify these unusually large or small values in our data. To calculate a z-score, recall that we take the individual value and subtract the mean and then divide this difference by the standard deviation.

$$z_i = \frac{x_i - \bar{x}}{S}$$

For any individual, the z-score tells us how many standard deviations the raw score for that individual deviates from the mean and in what direction. A positive z-score indicates the individual is above average and a negative z-score indicates the individual is below average.

Comments:

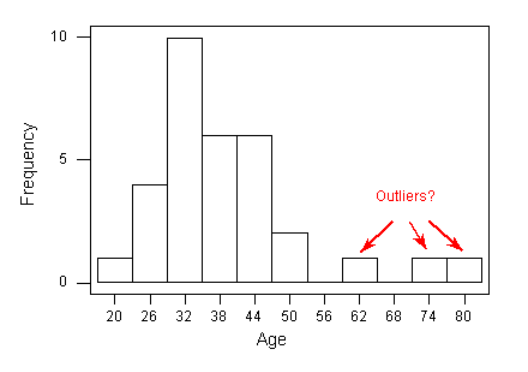
- Standardized scores can be used to help identify potential outliers
  - For approximately normal distributions, z-scores greater than 2 or less than -2 are rare (will happen approximately 5% of the time).
  - For any distribution, z-scores greater than 4 or less than -4 are rare (will happen less than 6.25% of the time).
- Standardized scores, along with other measures of position, are useful when comparing individuals in different datasets since the comparison takes into account the relative position of the individuals in their dataset. With z-scores, we can tell which individual has a relatively higher or lower position in their respective dataset.
- Later in the course, we will see that this idea of standardizing is used often in statistical analyses.

EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example (Link to the Best Actress Oscar Winners data).

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33

In previous examples, we identified three observations as outliers, two of which were classified as extreme outliers (ages of 61, 74 and 80)



The mean of this sample is 38.5 and the standard deviation is 12.95.

- The z-score for the actress with age = 80 is

$$z = \frac{80 - 38.5}{12.95} = 3.20$$

Thus, among our female Oscar winners from our sample, this actress is 3.20 standard deviations older than average.

Did I Get This?: Z-Scores

Role-Type Classification

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Video

Video: Role-Type Classification (Two Parts; 9:46 total time)

While it is fundamentally important to know how to describe the distribution of a single variable, most studies pose research questions that involve exploring the relationship between **two** (or more) variables. These research questions are investigated using a sample from the population of interest.

Reading: Form a Research Question (short)

Here are a few examples of such research questions with the two variables highlighted:

EXAMPLES:

1. Is there a relationship between **gender** and **test scores** on a particular standardized test? Other ways of phrasing the same research question:
   - Is performance on the test related to gender?
   - Is there a gender effect on test scores?
   - Are there differences in test scores between males and females?
2. How is the **number of calories** in a hot dog related to (or affected by) the **type of hot dog** (beef, meat or poultry)? In other words, are there differences in the number of calories among the three types of hot dogs?
3. Is there a relationship between the **type of light** a baby sleeps with (no light, night-light, lamp) and whether or not the child develops **nearsightedness**?
4. Are the **smoking habits** of a person (yes, no) related to the person's **gender**?
5. How well can we predict a student's freshman year **GPA** from his/her **SAT score**?
6. What is the relationship between driver's **age** and sign legibility **distance** (the maximum distance at which the driver can read a sign)?
7. Is there a relationship between the **time** a person has practiced driving while having a learner's permit, and **whether or not this person passed the driving test**?
8. Can you predict a person's **favorite type of music** (classical, rock, jazz) based on his/her **IQ level**?

Role of a Variable in a Study

Learning Objectives

LO 4.19: For a data analysis situation involving two variables, identify the role of each variable in the scenario.

In most studies involving two variables, each of the variables has a role. We distinguish between:

- the **response** variable — the outcome of the study; and
- the **explanatory** variable — the variable that claims to explain, predict or affect the response.

As we mentioned earlier the variable we wish to predict is commonly called the **dependent variable**, the **outcome** variable, or the **response** variable. Any variable we are using to predict (or explain differences) in the outcome is commonly called an **explanatory variable**, an **independent variable**, a **predictor** variable, or a **covariate**.

Comment:

- Typically the **explanatory** variable is denoted by X, and the **response** variable by Y.

Now let's go back to some of the examples and classify the two relevant variables according to their roles in the study:

✔ EXAMPLE 1:

Is there a relationship between **gender** and **test scores** on a particular standardized test? Other ways of phrasing the same research question:

- Is performance on the test related to gender?
- Is there a gender effect on test scores?
- Are there differences in test scores between males and females?

We want to explore whether the outcome of the study — the score on a test — is affected by the test-taker's gender. Therefore:

Gender is the **explanatory** variable

Test score is the **response** variable

✔ EXAMPLE 3:

Is there a relationship between the **type of light** a baby sleeps with (no light, night-light, lamp) and whether or not the child develops **nearsightedness**?

In this study we explore whether the nearsightedness of a person can be explained by the type of light that person slept with as a baby. Therefore:

Light type is the **explanatory** variable

Nearsightedness is the **response** variable

✔ EXAMPLE 5:

How well can we predict a student's freshman year **GPA** from his/her **SAT score**?

Here we are examining whether a student's SAT score is a good predictor for the student's GPA freshman year. Therefore:

SAT score is the **explanatory** variable

GPA of freshman year is the **response** variable

✔ EXAMPLE 7:

Is there a relationship between the **time** a person has practiced driving while having a learner's permit, and **whether or not this person passed the driving test**?

Here we are examining whether a person's outcome on the driving test (pass/fail) can be explained by the length of time this person has practiced driving prior to the test. Therefore:

Time is the **explanatory** variable

Driving test outcome is the **response** variable

Now, using the same reasoning, the following exercise will help you to classify the two variables in the other examples.

Learn By Doing: Role Classification

Many Students Wonder: Role Classification

Question: Is the role classification of variables always clear? In other words, is it always clear which of the variables is the explanatory and which is the response?

Answer: No. There are studies in which the role classification is not really clear. This mainly happens in cases when both variables are categorical or both are quantitative. An example is a study that explores the relationship between students' SAT Math and SAT Verbal scores. In cases like this, any classification choice would be fine (as long as it is consistent throughout the analysis).

Role-Type Classification

Learning Objectives

LO 4.20: Classify a data analysis situation involving two variables according to the "role-type classification."

If we further classify each of the two relevant variables according to **type** (categorical or quantitative), we get the following 4 possibilities for **"role-type classification"**

1. Categorical explanatory and quantitative response (Case CQ)
2. Categorical explanatory and categorical response (Case CC)
3. Quantitative explanatory and quantitative response (Case QQ)
4. Quantitative explanatory and categorical response (Case QC)

This role-type classification can be summarized and easily visualized in the following table (note that the explanatory variable is always listed first):

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C → C	C → Q
	Quantitative	Q → C	Q → Q

This role-type classification serves as the infrastructure for this entire section. In each of the 4 cases, different statistical tools (displays and numerical measures) should be used in order to explore the relationship between the two variables.

This suggests the following important principle:

PRINCIPLE: When confronted with a research question that involves exploring the relationship between two variables, the first and most crucial step is to determine which of the 4 cases represents the data structure of the problem. In other words, the first step should be classifying the two relevant variables according to their role and type, and only then can we determine what statistical tools should be used to analyze them.

Now let's go back to our 8 examples and determine which of the 4 cases represents the data structure of each:

EXAMPLE 1:

Is there a relationship between **gender** and **test scores** on a particular standardized test? Other ways of phrasing the same research question:

- Is performance on the test related to gender?
- Is there a gender effect on test scores?
- Are there differences in test scores between males and females?

We want to explore whether the outcome of the study — the score on a test — is affected by the test-taker's gender.

Gender is the **explanatory** variable and it is **categorical**.

Test score is the **response** variable and it is **quantitative**.

Therefore this is an example of **case C→Q**.

✔ EXAMPLE 3:

Is there a relationship between the **type of light** a baby sleeps with (no light, night-light, lamp) and whether or not the child develops **nearsightedness**?

In this study we explore whether the nearsightedness of a person can be explained by the type of light that person slept with as a baby.

Light type is the **explanatory** variable and it is **categorical**.

Nearsightedness is the **response** variable and it is **categorical**.

Therefore this is an example of **case C→C**.

✔ EXAMPLE 5:

How well can we predict a student's freshman year **GPA** from his/her **SAT score**?

Here we are examining whether a student's SAT score is a good predictor for the student's GPA freshman year.

SAT score is the **explanatory** variable and it is **quantitative**.

GPA of freshman year is the **response** variable and it is **quantitative**.

Therefore this is an example of **case Q→Q**.

✔ EXAMPLE 7:

Is there a relationship between the **time** a person has practiced driving while having a learner's permit, and **whether or not this person passed the driving test**?

Here we are examining whether a person's outcome on the driving test (pass/fail) can be explained by the length of time this person has practiced driving prior to the test.

Time is the **explanatory** variable and it is **quantitative**.

Driving test outcome is the **response** variable and it is **categorical**.

Therefore this is an example of **case Q→C**.

Now you complete the rest…

Learn By Doing: Role-Type Classification

The remainder of this section on exploring relationships will be guided by this role-type classification. In the next three parts we will elaborate on cases C→Q, C→C, and Q→Q. More specifically, we will learn the appropriate statistical tools (visual display and numerical measures) that will allow us to explore the relationship between the two variables in each of the cases. Case Q→C will **not** be discussed in this course, and is typically covered in more advanced courses. The section will conclude with a discussion on causal relationships.

Did I Get This?: Role-Type Classification

Summary (Unit 1)

(Optional) Outside Reading: Look at the Data! (≈1200 words)

(Optional) Outside Reading: Creating Data Files (≈1200 words)

This summary provides a quick recap of the material in the Exploratory Data Analysis unit. Please note that this summary **does not provide complete coverage** of the material, only lists the main points.

- The purpose of exploratory data analysis (EDA) is to convert the available **data** from their raw form to an informative one, in which the main features of the data are illuminated.
- When performing EDA, we should always:
  - use **visual displays** (graphs or tables) plus **numerical measures**.
  - describe the **overall pattern** and mention any **striking deviations** from that pattern.
  - **interpret** the results we find **in context**.
- When examining the **distribution** of a single variable, we distinguish between a **categorical** variable and a **quantitative** variable.
- The distribution of a **categorical** variable is summarized using:
  - Display: pie-chart or bar-chart (variation: pictogram → can be misleading — beware!)
  - Numerical measures: category (group) percentages.
- The distribution of a **quantitative** variable is summarized using:
  - Display: histogram (or stemplot, mainly for small data sets). When describing the distribution as displayed by the histogram, we should describe the:
    - Overall pattern → shape, center, spread.
    - Deviations from the pattern → outliers.
  - Numerical measures: descriptive statistics (measure of center plus measure of spread):
    - If distribution is symmetric with no outliers, use mean and standard deviation.
    - Otherwise, use the five-number summary, in particular, median and IQR (inter-quartile range).
- The five-number summary and the 1.5(IQR) Criterion for detecting outliers are the ingredients we need to build the **boxplot**. Boxplots are most effective when used side-by-side for comparing distributions (see also case C→Q in examining relationships).
- In the special case of a distribution having the normal shape, the Standard Deviation Rule applies. This rule tells us approximately what percent of the observations fall within 1,2, or 3 standard deviations away from the mean. In particular, when a distribution is approximately normal, almost all the observations (99.7%) fall within 3 standard deviations of the mean.
- When examining the relationship between two variables, the first step is to classify the two relevant variables according to their role and type:

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C→C	C→Q
	Quantitative	Q→C	Q→Q

and only then to determine the appropriate tools for summarizing the data. (We don't deal with case Q→C in this course).

- Case C→Q: Exploring the relationship amounts to **comparing the distributions** of the quantitative response variable for each category of the explanatory variable. To do this, we use:

- Display: side-by-side boxplots.
- Numerical measures: descriptive statistics of the response variable, for each value (category) of the explanatory variable separately.

- Case C→C: Exploring the relationship amounts to **comparing the distributions** of the categorical response variable, for each category of the explanatory variable. To do this, we use:
  - Display: two-way table.
  - Numerical measures: conditional percentages (of the response variable for each value (category) of the explanatory variable separately).
- Case Q→Q: We examine the relationship using:
  - Display: scatterplot. When describing the relationship as displayed by the scatterplot, be sure to consider:
    - Overall pattern → direction, form, strength.
    - Deviations from the pattern → outliers.

    Labeling the scatterplot (including a relevant third categorical variable in our analysis), might add some insight into the nature of the relationship.

  In the **special case** that the scatterplot displays a **linear** relationship (and only then), we supplement the scatterplot with:

  - **Numerical measures:** Pearson's correlation coefficient (r) **measures** the direction and, more importantly, the **strength of the linear relationship**. The closer r is to 1 (or -1), the stronger the positive (or negative) linear relationship. r is unitless, influenced by outliers, and should be used only as a supplement to the scatterplot.
  - When the relationship is linear (as displayed by the scatterplot, and supported by the correlation r), we can summarize the linear pattern using the **least squares regression line**. Remember that:
    - The slope of the regression line tells us the average change in the response variable that results from a 1-unit increase in the explanatory variable.
    - When using the regression line for predictions, you should beware of extrapolation.
- When examining the relationship between two variables (regardless of the case), any **observed relationship** (association) **does not imply causation**, due to the possible presence of lurking variables.
- When we include a lurking variable in our analysis, we might need to rethink the direction of the relationship → **Simpson's paradox**.

CHAPTER OVERVIEW

Unit 2: Producing Data

CO-1: Describe the roles biostatistics serves in the discipline of public health.

Video

Video: Producing Data Introduction (4:35)

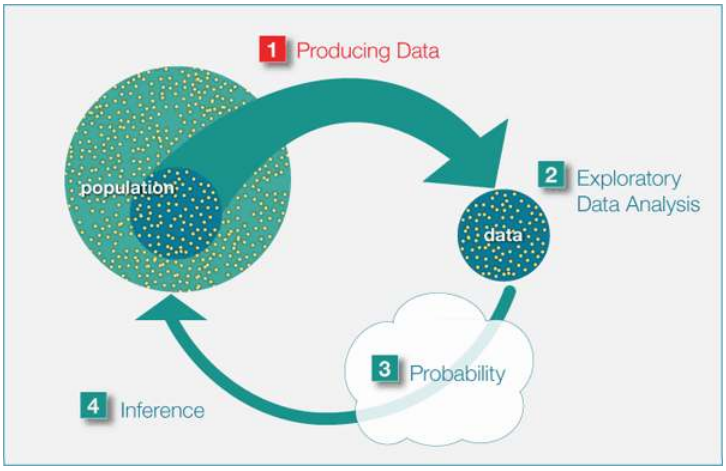
Review of the Big Picture

Learning Objectives

LO 1.3: Identify and differentiate between the components of the Big Picture of Statistics

Recall "The Big Picture," the four-step process that encompasses statistics: data production, exploratory data analysis, probability, and inference.

In the previous unit, we considered exploratory data analysis — the discovery of patterns in the raw data. In this unit, we go back and examine the first step in the process: the production of data. This unit has two main topics; **sampling** and **study design**.



Introduction to Producing Data

In the first step of the statistics "Big Picture," we produce data. The production of data has two stages.

- First we need to choose the individuals from the population that will be included in the sample.
- Then, once we have chosen the individuals, we need to collect data from them.

The first stage is called **sampling**, and the second stage is called **study design**.

As we have seen, exploratory data analysis seeks to illuminate patterns in the data by summarizing the distributions of quantitative or categorical variables, or the relationships between variables.

In the final part of the course, statistical inference, we will use the summaries about variables or relationships that were obtained in the study to draw conclusions about what is true for the entire population from which the sample was chosen.

For this process to "work" reliably, it is essential that the **sample** be truly **representative** of the larger population. For example, if researchers want to determine whether the antidepressant Zoloft is effective for teenagers in general, then it would not be a good idea to only test it on a sample of teens who have been admitted to a psychiatric hospital, because their depression may be more severe, and less treatable, than that of teens in general.

Thus, the very first stage in data production, **sampling**, must be carried out in such a way that the sample really does represent the population of interest.

Choosing a sample is only the first stage in producing data, so it is not enough to just make sure that the sample is representative. We must also remember that our summaries of variables and their relationships are only valid if these have been assessed properly.

For instance, if researchers want to test the effectiveness of Zoloft versus Prozac for treating teenagers, it would not be a good idea to simply compare levels of depression for a group of teenagers who happen to be using Zoloft to levels of depression for a group of teenagers who happen to be using Prozac. If they discover that one group of patients turns out to be less depressed, it could just be that teenagers with less serious depression are more likely to be prescribed one of the drugs over the other.

In situations like this, the **design** for producing data must be considered carefully. Studies should be designed to discover what we want to know about the variables of interest for the individuals in the sample.

In particular, if what you want to know about the variables is whether there is a causal relationship between them, special care should be given to the design of the study (since, as we know, association does not imply causation).

In this unit, we will focus on these two stages of data production: obtaining a sample, and designing a study.

Throughout this unit, we establish guidelines for the ideal production of data. While we will hold these guidelines as standards to strive for, realistically it is rarely possible to carry out a study that is completely free of flaws. Common sense must frequently be applied in order to decide which imperfections we can live with and which ones could completely undermine a study's results.

A sample that produces data that is not representative because of the systematic under- or over-estimation of the values of the variable of interest is called **biased**. Bias may result from either a poor sampling plan or from a poor design for evaluating the variable of interest.

We begin this unit by focusing on what constitutes a good — or a bad — sampling plan after which we will discuss study design.

Causation and Experiments

CO-3: Describe the strengths and limitations of designed experiments and observational studies.

Learning Objectives

LO 3.2: Explain how the study design impacts the types of conclusions that can be drawn.

Learning Objectives

LO 3.3: Identify and define key features of experimental design (randomized, blind etc.).

Video

Video: Causation and Experiments (8:57)

Recall that in an experiment, it is the researchers who assign values of the explanatory variable to the participants. The key to ensuring that individuals differ only with respect to explanatory values — which is also the key to establishing causation — lies in the way this assignment is carried out. Let's return to the smoking cessation study as a context to explore the essential ingredients of experimental design.

EXAMPLE:

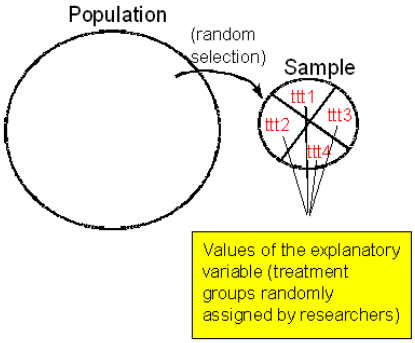
In our discussion of the distinction between observational studies and experiments, we described the following experiment: collect a representative sample of 1,000 individuals from the population of smokers who are just now trying to quit. We divide the sample into 4 groups of 250 and instruct each group to use a different method to quit. One year later, we contact the same 1,000 individuals and determine whose attempts succeeded while using our designated method.

This was an experiment, because the researchers themselves determined the values of the explanatory variable of interest for the individuals studied, rather than letting them choose.

We will begin by using the context of this smoking cessation example to illustrate the specialized vocabulary of experiments. First of all, the explanatory variable, or **factor**, in this case is the method used to quit. The different imposed values of the explanatory variable, or **treatments** (common abbreviation: ttt), consist of the four possible quitting methods. The groups receiving different treatments are called **treatment groups**. The group that tries to quit without drugs or therapy could be called the **control group** — those individuals on whom no specific treatment was imposed. Ideally, the **subjects** (human participants in an experiment) in each treatment group differ from those in the other treatment groups only with respect to the treatment (quitting method). As mentioned in our discussion of why lurking variables prevent us from establishing causation in observational studies, eliminating all other differences among treatment groups will be the key to asserting causation via an experiment. How can this be accomplished?

Randomized Controlled Experiments

Your intuition may already tell you, correctly, that **random assignment to treatments** is the best way to prevent treatment groups of individuals from differing from each other in ways other than the treatment assigned. Either computer software or tables can be utilized to accomplish the random assignment. The resulting design is called a **randomized controlled experiment,** because researchers control values of the explanatory variable with a randomization procedure. Under random assignment, the groups should not differ significantly with respect to any potential lurking variable. Then, if we see a relationship between the explanatory and response variables, we have evidence that it is a causal one.



Comments:

- Note that in a randomized controlled experiment, a randomization procedure may be used in two phases. First, a sample of subjects is collected. Ideally it would be a **random sample** so that it would be perfectly representative of the entire population.
- Often researchers have no choice but to recruit volunteers. Using volunteers may help to offset one of the drawbacks to experimentation which will be discussed later, namely the problem of noncompliance.
- Second, we **assign individuals randomly** to the treatment groups to ensure that the only difference between them will be due to the treatment and we can get evidence of causation. At this stage, randomization is vital.

Let's discuss some other issues related to experimentation.

Inclusion of a Control Group

A common misconception is that an experiment must include a control group of individuals receiving no treatment. There may be situations where a complete lack of treatment is not an option, or where including a control group is ethically questionable, or where researchers explore the effects of a treatment without making a comparison. Here are a few examples:

✔ EXAMPLE:

If doctors want to conduct an experiment to determine whether Prograf or Cyclosporin is more effective as an immunosuppressant, they could randomly assign transplant patients to take one or the other of the drugs. It would, of course, be unethical to include a control group of patients not receiving any immunosuppressants.

✔ EXAMPLE:

Recently, experiments have been conducted in which the treatment is a highly invasive brain surgery. The only way to have a legitimate control group in this case is to randomly assign half of the subjects to undergo the entire surgery except for the actual treatment component (inserting stem cells into the brain). This, of course, is also ethically problematic (but, believe it or not, is being done).

✔ EXAMPLE:

There may even be an experiment designed with only a single treatment. For example, makers of a new hair product may ask a sample of individuals to treat their hair with that product over a period of several weeks, then assess how manageable their hair has become. Such a design is clearly flawed because of the absence of a comparison group, but it is still an experiment because use of the product has been imposed by its manufacturers, rather than chosen naturally by the individuals. A flawed experiment is nevertheless an experiment.

Comment:

- The word **control** is used in at least three different senses.
  - In the context of observational studies, we **control for a confounding variable** by separating it out.
  - Referring to an experiment as a **controlled experiment** stresses that the values of the experiment's explanatory variables (factors) have been assigned by researchers, as opposed to having occurred naturally.

- In the context of experiments, the **control group** consists of subjects who do not receive a treatment, but who are otherwise handled identically to those who do receive the treatment.

Learn By Doing: Random Assignment to Treatment Groups (Software)

Blind and Double-Blind Experiments

Suppose the experiment about methods for quitting smoking were carried out with randomized assignments of subjects to the four treatments, and researchers determined that the percentage succeeding with the combination drug/therapy method was highest, and the percentage succeeding with no drugs or therapy was lowest. In other words, suppose there is clear evidence of an association between method used and success rate. Could it be concluded that the drug/therapy method causes success more than trying to quit without using drugs or therapy? Perhaps.

Although randomized controlled experiments do give us a better chance of pinning down the effects of the explanatory variable of interest, they are not completely problem-free. For example, suppose that the manufacturers of the smoking cessation drug had just launched a very high-profile advertising campaign with the goal of convincing people that their drug is extremely effective as a method of quitting.

Even with a randomized assignment to treatments, there would be an important difference among subjects in the four groups: those in the drug and combination drug/therapy groups would perceive their treatment as being a promising one, and may be more likely to succeed just because of added confidence in the success of their assigned method. Therefore, the ideal circumstance is for the subjects to be unaware of which treatment is being administered to them: in other words, subjects in an experiment should be (if possible) **blind** to which treatment they received.

How could researchers arrange for subjects to be blind when the treatment involved is a drug? They could administer a **placebo** pill to the control group, so that there are no psychological differences between those who receive the drug and those who do not. The word “placebo” is derived from a Latin word that means “to please.” It is so named because of the natural tendency of human subjects to improve just because of the “pleasing” idea of being treated, regardless of the benefits of the treatment itself. When patients improve because they are told they are receiving treatment, even though they are not actually receiving treatment, this is known as the **placebo effect.**

Next, how could researchers arrange for subjects to be blind when the treatment involved is a type of therapy? This is more problematic. Clearly, subjects must be aware of whether they are undergoing some type of therapy or not. There is no practical way to administer a “placebo” therapy to some subjects. Thus, the relative success of the drug/therapy treatment may be due to subjects’ enhanced confidence in the success of the method they happened to be assigned. We may feel fairly certain that the method itself causes success in quitting, but we cannot be absolutely sure.

When the response of interest is fairly straightforward, such as giving up cigarettes or not, then recording its values is a simple process in which researchers need not use their own judgment in making an assessment. There are many experiments where the response of interest is less definite, such as whether or not a cancer patient has improved, or whether or not a psychiatric patient is less depressed. In such cases, it is important for researchers who evaluate the response to be **blind** to which treatment the subject received, in order to prevent the **experimenter effect** from influencing their assessments. If neither the subjects nor the researchers know who was assigned what treatment, then the experiment is called **double-blind.**

The most reliable way to determine whether the explanatory variable is actually causing changes in the response variable is to carry out a **randomized controlled double-blind experiment**. Depending on the variables of interest, such a design may not be entirely feasible, but the closer researchers get to achieving this ideal design, the more convincing their claims of causation (or lack thereof) are.

Did I Get This?: Experiments

Pitfalls in Experimentation

Some of the inherent difficulties that may be encountered in experimentation are the Hawthorne effect, lack of realism, noncompliance, and treatments that are unethical, impossible, or impractical to impose.

We already introduced a hypothetical experiment to determine if people tend to snack more while they watch TV:

- Recruit participants for the study.
- While they are presumably waiting to be interviewed, half of the individuals sit in a waiting room with snacks available and a TV on. The other half sit in a waiting room with snacks available and no TV, just magazines.
- Researchers determine whether people consume more snacks in the TV setting.

Suppose that, in fact, the subjects who sat in the waiting room with the TV consumed more snacks than those who sat in the room without the TV. Could we conclude that in their everyday lives, and in their own homes, people eat more snacks when the TV is on? Not necessarily, because people's behavior in this very controlled setting may be quite different from their ordinary behavior.

If they suspect their snacking behavior is being observed, they may alter their behavior, either consciously or subconsciously. This phenomenon, whereby people in an experiment behave differently from how they would normally behave, is called the **Hawthorne effect**. Even if they don't suspect they are being observed in the waiting room, the relationship between TV and snacking in the waiting room might not be representative of what it is in real life.

One of the greatest advantages of an experiment — that researchers take control of the explanatory variable — can also be a disadvantage in that it may result in a rather unrealistic setting. **Lack of realism** (also called **lack of ecological validity**) is a possible drawback to the use of an experiment rather than an observational study to explore a relationship. Depending on the explanatory variable of interest, it may be quite easy or it may be virtually impossible to take control of the variable's values and still maintain a fairly natural setting.

In our hypothetical smoking cessation example, both the observational study and the experiment were carried out on a random sample of 1,000 smokers with intentions to quit. In the case of the observational study, it would be reasonably feasible to locate 1,000 such people in the population at large, identify their intended method, and contact them again a year later to establish whether they succeeded or not.

In the case of the experiment, it is not so easy to take control of the explanatory variable (cessation method) merely by telling all 1,000 subjects what method they must use. **Noncompliance** (failure to submit to the assigned treatment) could enter in on such a large scale as to render the results invalid.

In order to ensure that the subjects in each treatment group actually undergo the assigned treatment, researchers would need to pay for the treatment and make it easily available. The cost of doing that for a group of 1,000 people would go beyond the budget of most researchers.

Even if the drugs or therapy were paid for, it is very unlikely that most of the subjects contacted at random would be willing to use a method not of their own choosing, but dictated by the researchers. From a practical standpoint, such a study would most likely be carried out on a smaller group of volunteers, recruited via flyers or some other sort of advertisement.

The fact that they are volunteers might make them somewhat different from the larger population of smokers with intentions to quit, but it would reduce the more worrisome problem of non-compliance. Volunteers may have a better overall chance of success, but if researchers are primarily concerned with which method is most successful, then the relative success of the various methods should be roughly the same for the volunteer sample as it would be for the general population, as long as the methods are randomly assigned. Thus, the most vital stage for randomization in an experiment is during the assignment of treatments, rather than the selection of subjects.

There are other, more serious drawbacks to experimentation, as illustrated in the following hypothetical examples:

✔ EXAMPLE:

Suppose researchers want to determine if the drug Ecstasy causes memory loss. One possible design would be to take a group of volunteers and randomly assign some to take Ecstasy on a regular basis, while the others are given a placebo. Test them periodically to see if the Ecstasy group experiences more memory problems than the placebo group.

The obvious flaw in this experiment is that it is unethical (and actually also illegal) to administer a dangerous drug like Ecstasy, even if the subjects are volunteers. The only feasible design to seek answers to this particular research question would be an observational study.

EXAMPLE:

Suppose researchers want to determine whether females wash their hair more frequently than males.

It is impossible to assign some subjects to be female and others male, and so an experiment is not an option here. Again, an observational study would be the only way to proceed.

EXAMPLE:

Suppose researchers want to determine whether being in a lower income bracket may be responsible for obesity in women, at least to some extent, because they can't afford more nutritious meals and don't have the means to participate in fitness activities.

The socioeconomic status of the study subject is a variable that cannot be controlled by the researchers, so an experiment is impossible. (Even if the researchers could somehow raise the money to provide a random sample of women with substantial salaries, the effects of their eating habits during their lives before the study began would still be present, and would affect the study's outcome.)

These examples should convince you that, depending on the variables of interest, researching their relationship via an experiment may be too unrealistic, unethical, or impractical. Observational studies are subject to flaws, but often they are the only recourse.

Let's summarize what we've learned so far:

1. Observational studies:

- The explanatory variable's values are allowed to occur naturally.
- Because of the possibility of lurking variables, it is difficult to establish causation.
- If possible, control for suspected lurking variables by studying groups of similar individuals separately.
- Some lurking variables are difficult to control for; others may not be identified.

2. Experiments

- The explanatory variable's values are controlled by researchers (treatment is imposed).
- Randomized assignment to treatments automatically controls for all lurking variables.
- Making subjects blind avoids the placebo effect.
- Making researchers blind avoids conscious or subconscious influences on their subjective assessment of responses.
- A randomized controlled double-blind experiment is generally optimal for establishing causation.
- A lack of realism may prevent researchers from generalizing experimental results to real-life situations.
- Noncompliance may undermine an experiment. A volunteer sample might solve (at least partially) this problem.
- It is impossible, impractical, or unethical to impose some treatments.

More About Experiments

CO-3: Describe the strengths and limitations of designed experiments and observational studies.

Learning Objectives

LO 3.2: Explain how the study design impacts the types of conclusions that can be drawn.

Learning Objectives

LO 3.3: Identify and define key features of experimental design (randomized, blind etc.).

Video

Video: More About Experiments (4:09)

Experiments With More Than One Explanatory Variable

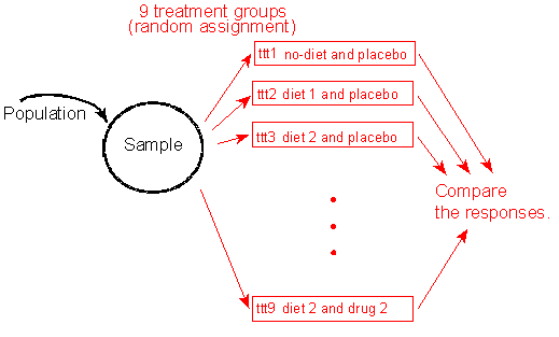
It is not uncommon for experiments to feature two or more explanatory variables (called factors). In this course, we focus on exploratory data analysis and statistical inference in situations which involve only one explanatory variable. Nevertheless, we will now consider the design for experiments involving several explanatory variables, in order to familiarize students with their basic structure.

EXAMPLE:

Suppose researchers are not only interested in the effect of diet on blood pressure, but also the effect of two new drugs. Subjects are assigned to either Control Diet (no restrictions), Diet #1, or Diet #2, (the variable diet has, then, 3 possible values) and are also assigned to receive either Placebo, Drug #1, or Drug #2 (the variable Drug, then, also has three values). This is an example where the experiment has two explanatory variables and a response variable. In order to set up such an experiment, there has to be **one treatment group for every combination of categories of the two explanatory variables**. Thus, in this case there are 3 * 3 = 9 combinations of the two variables to which the subjects are assigned. The treatment groups are illustrated and labeled in the following table:

	No-diet	Special diet1	Special diet 2
Placebo	ttt1	ttt2	ttt3
Drug 1	ttt4	ttt5	ttt6
Drug 2	ttt7	ttt8	ttt9

Subjects would be randomly assigned to one of the nine treatment groups. If we find differences in the proportions of subjects who achieve the lower “moderate zone” blood pressure among the nine treatment groups, then we have evidence that the diets and/or drugs may be effective for reducing blood pressure.



Comments:

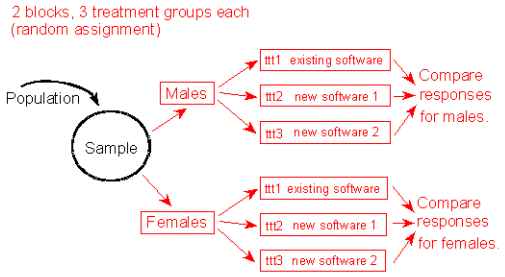
- Recall that randomization may be employed at two stages of an experiment: in the selection of subjects, and in the assignment of treatments. The former may be helpful in allowing us to generalize what occurs among our subjects to what would occur in the general population, but the reality of most experimental settings is that a convenience or volunteer sample is used. Most likely the blood pressure study described above would use volunteer subjects. The important thing is to make sure these subjects are randomly assigned to one of the nine treatment combinations.
- In order to gain optimal information about individuals in all the various treatment groups, we would like to make assignments not just randomly, but also evenly. If there are 90 subjects in the blood pressure study described above, and 9 possible treatment groups, then each group should be filled randomly with 10 individuals. A simple random sample of 10 could be taken from the larger group of 90, and those individuals would be assigned to the first treatment group. Next, the second treatment group would be filled by a simple random sample of 10 taken from the remaining 80 subjects. This process would be repeated until all 9 groups are filled with 10 individuals each.

Did I Get This?: Experiments #2

Modifications to Randomization

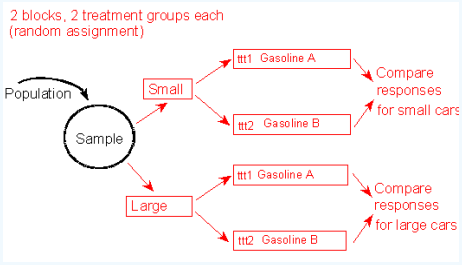
In some cases, an experiment's design may be enhanced by relaxing the requirement of total randomization and **blocking** the subjects first, dividing them into groups of individuals who are similar with respect to an outside variable that may be important in the relationship being studied. This can help ensure that the effect of treatments, as well as background variables, are most precisely measured. In blocking, we simply split the sampled subjects into blocks based upon the different values of the background variable, and then randomly allocate treatments within each block. Thus, blocking in the assignment of subjects is analogous to stratification in sampling.

For example, consider again our experiment examining the differences between three versions of software from the last Learn By Doing activity. If we suspected that gender might affect individuals' software preferences, we might choose to allocate subjects to separate blocks, one for males and one for females. Within each block, subjects are randomly assigned to treatments and the treatment proceeds as usual. A diagram of blocking in this situation is below:



EXAMPLE:

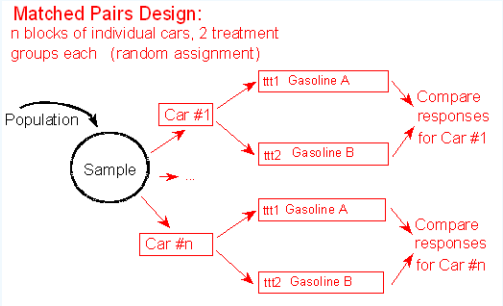
Suppose producers of gasoline want to compare which of two types of gas results in better mileage for automobiles. In case the size of the vehicle plays a role in the effectiveness of different types of gasoline, they could first block by vehicle size, then randomly assign some cars within each block to Gasoline A and others to Gasoline B:



In the extreme, researchers may examine a relationship for a sample of blocks of just two individuals who are similar in many important respects, or even the same individual whose responses are compared for two explanatory values.

EXAMPLE:

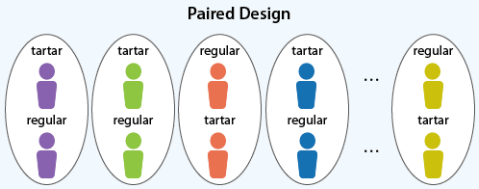
For example, researchers could compare the effects of Gasoline A and Gasoline B when both are used on the same car, for a sample of many cars of various sizes and models.



Such a study design, called matched pairs, may enable us to pinpoint the effects of the explanatory variable by comparing responses for the same individual under two explanatory values, or for two individuals who are as similar as possible except that the first gets one treatment, and the second gets another (or serves as the control). Treatments should usually be assigned at random within each pair, or the order of treatments should be randomized for each individual. In our gasoline example, for each car the order of testing (Gasoline A first, or Gasoline B first) should be randomized.

EXAMPLE:

Suppose researchers want to compare the relative merits of toothpastes with and without tartar control ingredients. In order to make the comparison between individuals who are as similar as possible with respect to background and diet, they could obtain a sample of identical twins. One of each pair would randomly be assigned to brush with the tartar control toothpaste, while the other would brush with regular toothpaste of the same brand. These would be provided in unmarked tubes, so that the subjects would be blind. To make the experiment double-blind, dentists who evaluate the results would not know who used which toothpaste.



"Before-and-after" studies are another common type of matched pairs design. For each individual, the response variable of interest is measured twice: first before the treatment, then again after the treatment. The categorical explanatory variable is which treatment was applied, or whether a treatment was applied, to that participant.

Comment:

- We have explained data production as a two-stage process: first obtain the sample, then evaluate the variables of interest via an appropriate study design. Even though the steps are carried out in this order chronologically, it is generally best for researchers to decide on a study design before they actually obtain the sample. For the toothpaste example above, researchers would first decide to use the matched pairs design, then obtain a sample of identical twins, then carry out the experiment and assess the results.

These examples should convince you that, depending on the variables of interest, researching their relationship via an experiment may be too unrealistic, unethical, or impractical. Observational studies are subject to flaws, but often they are the only recourse.

Did I Get This?: More About Experiments

Causation and Observational Studies

CO-3: Describe the strengths and limitations of designed experiments and observational studies.

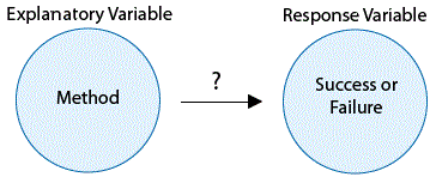
Learning Objectives

LO 3.2: Explain how the study design impacts the types of conclusions that can be drawn.

Video

Video: Causation and Observational Studies (3:09)

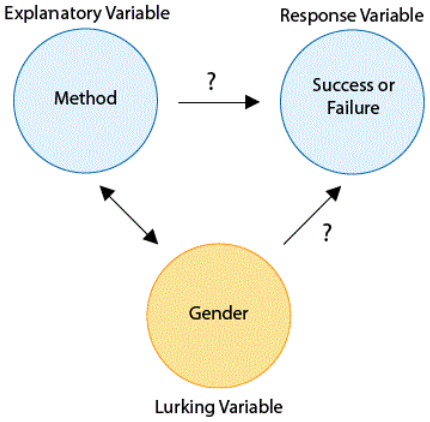
Suppose the **observational study** described earlier was carried out, and researchers determined that the percentage succeeding with the combination drug/therapy method was highest, while the percentage succeeding with neither therapy nor drugs was lowest. In other words, suppose there is clear evidence of an association between method used and success rate. Could they then conclude that the combination drug/therapy method causes success more than using neither therapy nor a drug?



It is at precisely this point that we confront the underlying weakness of most observational studies: some members of the sample have opted for certain values of the explanatory variable (method of quitting), while others have opted for other values. It could be that those individuals may be different in additional ways that would also play a role in the response of interest.

For instance, suppose women are more likely to choose certain methods to quit, and suppose women in general tend to quit more successfully than men. The data would make it appear that the method itself was responsible for success, whereas in truth it may just be that being female is the reason for success.

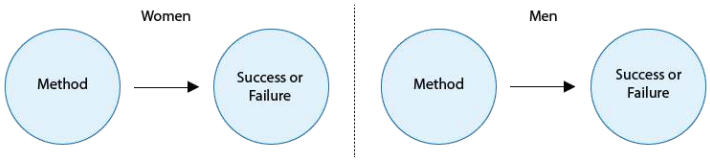
We can express this scenario in terms of the key variables involved. In addition to the explanatory variable (method) and the response variable (success or failure), a third, **lurking** variable (gender) is tied in (or **confounded**) with the explanatory variable's values, and may itself cause the response to be a success or failure. The following diagram illustrates this situation.



Since the difficulty arises because of the lurking variable's values being tied in with those of the explanatory variable, one way to attempt to unravel the true nature of the relationship between explanatory and response variables is to separate out the effects of the lurking variable. In general, we **control** for the effects of a lurking variable by separately studying groups that are defined by this variable.

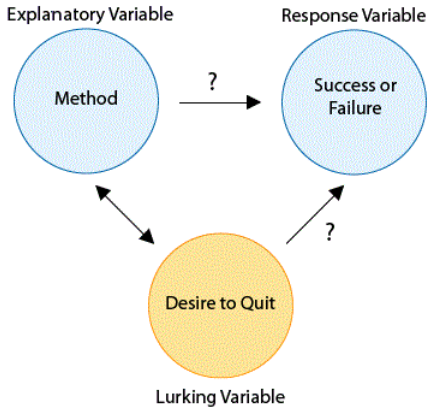
Caution

We could control for the lurking variable "gender" by **studying women and men separately**. Then, if both women and men who chose one method have higher success rates than those opting for another method, we would be closer to producing evidence of causation.



The diagram above demonstrates how straightforward it is to control for the lurking variable gender.

Notice that we did not claim that controlling for gender would allow us to make a definite claim of causation, only that we would be closer to establishing a causal connection. This is due to the fact that other lurking variables may also be involved, such as the level of the participants' desire to quit. Specifically, those who have chosen to use the drug/therapy method may already be the ones who are most determined to succeed, while those who have chosen to quit without investing in drugs or therapy may, from the outset, be less committed to quitting. The following diagram illustrates this scenario.



To attempt to control for this lurking variable, we could interview the individuals at the outset in order to rate their desire to quit on a scale of 1 (weakest) to 5 (strongest), and study the relationship between method and success separately for each of the five groups. But desire to quit is obviously a very subjective thing, difficult to assign a specific number to. Realistically, we may be unable to effectively control for the lurking variable "desire to quit."

Furthermore, who's to say that gender and/or desire to quit are the only lurking variables involved? There may be other subtle differences among individuals who choose one of the four various methods that researchers fail to imagine as they attempt to control for possible lurking variables.

For example, smokers who opt to quit using neither therapy nor drugs may tend to be in a lower income bracket than those who opt for (and can afford) drugs and/or therapy. Perhaps smokers in a lower income bracket also tend to be less successful in quitting because more of their family members and co-workers smoke. Thus, socioeconomic status is yet another possible lurking variable in the relationship between cessation method and success rate.

It is because of the existence of a virtually unlimited number of potential lurking variables that we can never be 100% certain of a claim of causation based on an observational study. On the other hand, observational studies are an extremely common tool used by researchers to attempt to draw conclusions about causal connections.

If great care is taken to control for the most likely lurking variables (and to avoid other pitfalls which we will discuss presently), and if common sense indicates that there is good reason for one variable to cause changes in the other, then researchers may assert that an observational study provides good evidence of causation.

Observational studies are subject to other pitfalls besides lurking variables, arising from various aspects of the design for evaluating the explanatory and response values. The next pair of examples illustrates some other difficulties that may arise.

EXAMPLE:

Suppose researchers want to determine if people tend to snack more while they watch TV. One possible design that we considered was to recruit participants for an observational study, and give them journals to record their hourly activities the following day, including TV watched and snacks consumed. Then they could review the journals to determine if snack consumption was higher during TV times.

We identified this as a prospective observational study, carried forward in time. Studying people in the more natural setting of their own homes makes the study more realistic than a contrived experimental setting. Still, when people are obliged to record their behavior as it occurs, they may be too self-conscious to act naturally. They may want to avoid embarrassment and so they may cut back on their TV viewing, or their snack consumption, or the combination of the two.

Yet another possible design is to recruit participants for a retrospective observational study. Ask them to recall, for each hour of the previous day, whether they were watching TV, and what snacks they consumed each hour. Determine if food consumption was higher during the TV times.

This design has the advantage of not disturbing people's natural behavior in terms of TV viewing or snacking. It has the disadvantage of relying on people's memories to record those variables' values from the day before. But one day is a relatively short period of time to remember such details, and as long as people are willing to be honest, the results of this study could be fairly reliable. The issue of eliciting honest responses will be addressed in our discussion of sample surveys.

By now you should have an idea of how **difficult — or perhaps even impossible —** it is **to establish causation in an observational study,** especially due to the problem of lurking variables.

The key to establishing causation is to rule out the possibility of any lurking variable, or in other words, to ensure that individuals differ **only with respect to the values of the explanatory variable**.

In general, this is a goal which we have a much better chance of accomplishing by carrying out a well-designed experiment.

Designing Studies

CO-3: Describe the strengths and limitations of designed experiments and observational studies.

Video

Video: Designing Studies (1:34)

Now that we have learned about the first stage of data production — sampling — we can move on to the next stage — designing studies.

Introduction

Obviously, sampling is not done for its own sake. After this first stage in the data production process is completed, we come to the second stage, that of gaining information about the variables of interest from the sampled individuals. Now we'll discuss three study designs; each design enables you to determine the values of the variables in a different way.

You can:

- Carry out an **observational study**, in which values of the variable or variables of interest are recorded as they naturally occur. There is no interference by the researchers who conduct the study.
- Take a **sample survey**, which is a particular type of observational study in which individuals report variables' values themselves, frequently by giving their opinions.
- Perform an **experiment**. Instead of assessing the values of the variables as they naturally occur, the researchers interfere, and they are the ones who assign the values of the explanatory variable to the individuals. The researchers "take control" of the values of the explanatory variable because they want to see how changes in the value of the explanatory variable affect the response variable. (Note: By nature, any experiment involves at least two variables.)

The type of design used, and the details of the design, are crucial, since they will determine what kind of conclusions we may draw from the results. In particular, when studying relationships in the Exploratory Data Analysis unit, we stressed that an association between two variables does not guarantee that a causal relationship exists. Here we will explore how the details of a study design play a crucial role in determining our ability to establish evidence of causation.

Here is how this topic is organized:

We'll start by learning how to identify study types. In particular, we will highlight the distinction between observational studies and experiments.

We will then discuss each of the three study designs mentioned above.

- We'll discuss observational studies, focusing on why it is difficult to establish causation in these type of studies, as well as other possible flaws.
- We'll then focus on experiments, learning, among other things, that when appropriately designed, experiments **can** provide evidence of causation.
- We'll end by discussing surveys and sample size

Identifying Study Design

Learning Objectives

LO 3.1: Identify the design of a study (controlled experiment vs. observational study)

Because each type of study design has its own advantages and trouble spots, it is important to begin by determining what type of study we are dealing with. The following example helps to illustrate how we can distinguish among the three basic types of design

mentioned in the introduction — observational studies, sample surveys, and experiments.

EXAMPLE:

Suppose researchers want to determine whether people tend to snack more while they watch television. In other words, the researchers would like to explore the relationship between the explanatory variable "TV" (a categorical variable that takes the values "on'" and "not on") and the response variable "snack consumption."

Identify each of the following designs as being an observational study, a sample survey, or an experiment.

1. Recruit participants for a study. While they are presumably waiting to be interviewed, half of the individuals sit in a waiting room with snacks available and a TV on. The other half sit in a waiting room with snacks available and no TV, just magazines. Researchers determine whether people consume more snacks in the TV setting.

   This is an **experiment**, because the researchers take control of the explanatory variable of interest (TV on or not) by **assigning** each individual to either watch TV or not, and determine the effect that has on the response variable of interest (snack consumption).

   ---

2. Recruit participants for a study. Give them journals to record hour by hour their activities the following day, including when they watch TV and when they consume snacks. Determine if snack consumption is higher during TV times.

   This is an **observational study**, because the participants themselves determine whether or not to watch TV. There is no attempt on the researchers' part to interfere.

   ---

3. Recruit participants for a study. Ask them to recall, for each hour of the previous day, whether they were watching TV, and what snacks they consumed each hour. Determine whether snack consumption was higher during the TV times.

   This is also an **observational study**; again, it was the participants themselves who decided whether or not to watch TV. Do you see the difference between 2 and 3? See the comment below.

   ---

4. Poll a sample of individuals with the following question: While watching TV, do you tend to snack: (a) less than usual; (b) more than or usual; or (c) the same amount as usual?

   This is a **sample survey**, because the individuals self-assess the relationship between TV watching and snacking.

Comment:

- Notice that in Example 2, the values of the variables of interest (TV watching and snack consumption) are recorded forward in time. Such observational studies are called **prospective**. In contrast, in Example 3, the values of the variables of interest are recorded backward in time. This is called a **retrospective** observational study.

Did I Get This?: Study Design

While some studies are designed to gather information about a single variable, many studies attempt to draw conclusions about the relationship between two variables. In particular, researchers often would like to produce evidence that one variable actually causes changes in the other.

For example, the research question addressed in the previous example sought to establish evidence that watching TV could cause an increase in snacking. Such studies may be especially useful and interesting, but they are also especially vulnerable to flaws that could invalidate the conclusion of causation.

In several of the examples we will see that although evidence of an association between two variables may be quite clear, the question of whether one variable is actually causing changes in the other may be too murky to be entirely resolved. In general, with a well-designed experiment we have a better chance of establishing causation than with an observational study.

However, experiments are also subject to certain pitfalls, and there are many situations in which an experiment is not an option. A well-designed observational study may still provide fairly convincing evidence of causation under the right circumstances.

Experiments vs. Observational Studies

Before assessing the effectiveness of observational studies and experiments for producing evidence of a causal relationship between two variables, we will illustrate the essential differences between these two designs.

EXAMPLE:

Every day, a huge number of people are engaged in a struggle whose outcome could literally affect the length and quality of their life: they are trying to quit smoking. Just the array of techniques, products, and promises available shows that quitting is not easy, nor is its success guaranteed. Researchers would like to determine which of the following is the best method:

1. Drugs that alleviate nicotine addiction.
2. Therapy that trains smokers to quit.
3. A combination of drugs and therapy.
4. Neither form of intervention (quitting "cold turkey").

The explanatory variable is the method (1, 2, 3 or 4), while the response variable is eventual success or failure in quitting. In an observational study, values of the explanatory variable occur naturally. In this case, this means that the participants themselves choose a method of trying to quit smoking. In an experiment, researchers assign the values of the explanatory variable. In other words, they tell people what method to use. Let us consider how we might compare the four techniques, via either an observational study or an experiment.

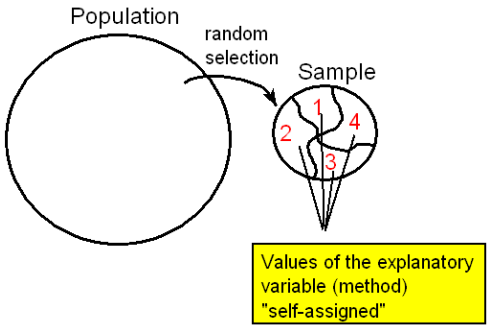
1. An **observational study** of the relationship between these two variables requires us to collect a representative sample from the population of smokers who are beginning to try to quit. We can imagine that a substantial proportion of that population is trying one of the four above methods. In order to obtain a representative sample, we might use a nationwide telephone survey to identify 1,000 smokers who are just beginning to quit smoking. We record which of the four methods the smokers use. One year later, we contact the same 1,000 individuals and determine whether they succeeded.

---

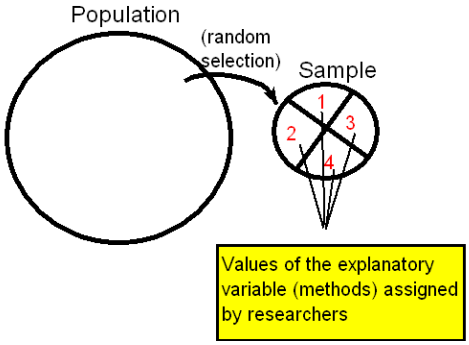
2. In an **experiment**, we again collect a representative sample from the population of smokers who are just now trying to quit, using a nationwide telephone survey of 1,000 individuals. This time, however, we divide the sample into 4 groups of 250 and **assign** each group to use one of the four methods to quit. One year later, we contact the same 1,000 individuals and determine whose attempts succeeded while using our designated method.

The following figures illustrate the two study designs:

Observational study:



Experiment:



Both the observational study and the experiment begin with a random sample from the population of smokers just now beginning to quit. In both cases, the individuals in the sample can be divided into categories based on the values of the explanatory variable: method used to quit. The response variable is success or failure after one year. Finally, in both cases, we would assess the relationship between the variables by comparing the proportions of success of the individuals using each method, using a two-way table and conditional percentages.

The only difference between the two methods is the way the sample is divided into categories for the explanatory variable (method). In the observational study, individuals are divided based upon the method by which they **choose** to quit smoking. The researcher does not assign the values of the explanatory variable, but rather records them as they naturally occur. In the experiment, the researcher **deliberately assigns** one of the four methods to each individual in the sample. The researcher intervenes by controlling the explanatory variable, and then assesses its relationship with the response variable.

Now that we have outlined two possible study designs, let's return to the original question: which of the four methods for quitting smoking is most successful? Suppose the study's results indicate that individuals who try to quit with the combination drug/therapy method have the highest rate of success, and those who try to quit with neither form of intervention have the lowest rate of success, as illustrated in the hypothetical two-way table below:

	Quit	Didn't Quit	Total	% Who Quit
Cold Turkey	12	238	250	5%
Drugs only	60	190	250	24%
Therapy only	59	191	250	24%
Drugs & Therapy	83	167	250	33%

Can we conclude that using the combination drugs and therapy method caused the smokers to quit most successfully? Which type of design was implemented will play an important role in the answer to this question.

> **Did I Get This?:** Study Design #2

Sample Surveys

CO-3: Describe the strengths and limitations of designed experiments and observational studies.

Learning Objectives

LO 3.2: Explain how the study design impacts the types of conclusions that can be drawn.

Learning Objectives

LO 3.4: Identify common problems with surveys and determine the potential impact(s) of each on the collected data and the accuracy of the data.

Video

Video: Sample Surveys (2:58)

Concepts of Sample Surveys

A sample survey is a particular type of observational study in which individuals report variables' values themselves, frequently by giving their opinions. Researchers have several options to choose from when deciding how to survey the individuals involved: in person, or via telephone, Internet, or mail.

The following issues in the design of sample surveys will be discussed:

- open vs. closed questions
- unbalanced response options
- leading questions
- planting ideas with questions
- complicated questions
- sensitive questions

These issues are best illustrated with a variety of concrete examples.

Suppose you want to determine the musical preferences of all students at your university, based on a sample of students. In the **Sampling** section, we discussed various ways to obtain the sample, such as taking a simple random sample from all students at the university, then contacting the chosen subjects via email to request their responses and following up with a second email to those who did not respond the first time.

This method would ensure a sample that is fairly representative of the entire population of students at the university, and avoids the bias that might result from a flawed design such as a convenience sample or a volunteer sample.

However, even if we managed to select a representative sample for a survey, we are not yet home free: we must still compose the survey question itself so that the information we gather from the sampled students correctly represents what is true about their musical preferences.

Let's consider some possibilities:

Question: "What is your favorite kind of music?"

This is what we call an **open question**, which allows for almost unlimited responses. It may be difficult to make sense of all the possible categories and subcategories of music that survey respondents could come up with.

Some may be more general than what you had in mind ("I like modern music the best") and others too specific ("I like Japanese alternative electronic rock by Cornelius"). Responses are much easier to handle if they come from a **closed question:**

> **Question:** Which of these types of music do you prefer: classical, rock, pop, or hip-hop?

What will happen if a respondent is asked the question as worded above, and he or she actually prefers jazz or folk music or gospel? He or she may pick a second-favorite from the options presented, or try to pencil in the real preference, or may just not respond at all. Whatever the outcome, it is likely that overall, the responses to the question posed in this way will not give us very accurate information about general music preferences. If a closed question is used, then great care should be taken to include all the reasonable options that are possible, including "not sure." Also, in case an option was overlooked, "other:_____" should be included for the sake of thoroughness.

Many surveys ask respondents to assign a rating to a variable, such as in the following:

> **Question:** How do you feel about classical music? Circle one of these: I love it, I like it very much, I like it, I don't like it, I hate it.

Notice that the options provided are rather "top-heavy," with three favorable options vs. two unfavorable. If someone feels somewhat neutral, they may opt for the middle choice, "I like it," and a summary of the survey's results would distort the respondents' true opinions.

Some survey questions are either deliberately or unintentionally biased towards certain responses:

> **Question:** "Do you agree that classical music is the best type of music, because it has survived for centuries and is not only enjoyable, but also intellectually rewarding? (Answer yes or no.)"

This sort of wording puts ideas in people's heads, urging them to report a particular opinion. One way to test for bias in a survey question is to ask yourself, "Just from reading the question, would a respondent have a good idea of what response the surveyor is hoping to elicit?" If the answer is yes, then the question should have been worded more neutrally.

Sometimes, survey questions are ordered in such a way as to deliberately bias the responses by planting an idea in an earlier question that will sway people's thoughts in a later question.

> **Question:** In the year 2002, there was much controversy over the fact that the Augusta National Golf Club, which hosts the Masters Golf Tournament each year, does not accept women as members. Defenders of the club created a survey that included the following statements. Respondents were supposed to indicate whether they agreed or disagreed with each statement:
>
> "The First Amendment of the U.S. Constitution applies to everyone regardless of gender, race, religion, age, profession, or point of view."
>
> "The First Amendment protects the right of individuals to create a private organization consisting of a specific group of people based on age, gender, race, ethnicity, or interest."
>
> "The First Amendment protects the right of organizations like the Boy Scouts, the Girls Scouts, and the National Association for the Advancement of Colored People to exist."
>
> "Individuals have a right to join a private group, club, or organization that consists of people who share the same interests and personal backgrounds as they do if they so desire."
>
> "Private organizations that are not funded by the government should be allowed to decide who becomes a member and who does not become a member on their own, without being forced to take input from other outside people or organizations."

Notice how the first and second statements steer people to favor the opinion that specialized groups may form private clubs. The third statement reminds people of organizations that are formed by groups on the basis of gender and race, setting the stage for them to agree with the fourth statement, which supports people's rights to join any private club. This in turn leads into the fifth statement, which focuses on a private organization's right to decide on its membership. As a group, the questions attempt to relentlessly steer a respondent towards ultimately agreeing with the club's right to exclude women.

Sometimes surveyors attempt to get feedback on more than one issue at a time.

Question: "Do you agree or disagree with this statement: 'I don't go out of my way to listen to modern music unless there are elements of jazz, or else lyrics that are clear and make sense.'"

Put yourself in the place of people who enjoy jazz and straightforward lyrics, but don't have an issue with music being "too modern," per se. The logic of the question (or lack thereof) may escape the respondents, and they would be too confused to supply an answer that correctly conveys their opinion. Clearly, simple questions are much better than complicated ones; rather than try to gauge opinions on several issues at once, complex survey questions like this should be broken down into shorter, more concise ones.

Depending on the topic, we cannot always assume that survey respondents will answer honestly.

Question1: "Have you eaten rutabagas in the past year?"

If respondents answer no, then we have good reason to believe that they did not eat rutabagas in the past year.

Question2: "Have you used illegal drugs in the past year?"

If respondents answer no, then it is still a possibility that they did use illegal drugs, but didn't want to admit it.

Effective techniques for collecting accurate data on sensitive questions are a main area of inquiry in statistics. One simple method is **randomized response**, which allows individuals in the sample to answer anonymously, while the researcher still gains information about the population. This technique is best illustrated by an example.

✔ EXAMPLE:

For the question, "Have you used illegal drugs in the past year?" respondents are told to flip a fair coin (in private) before answering and then answer based on the result of the coin flip: if the coin flip results in "Heads," they should answer "Yes" (regardless of the truth), if a coin flip results in "Tails," they should answer truthfully. Thus, roughly half of the respondents are "truth-tellers," and the other half give the uncomfortable answer "Yes," without the interviewer's knowledge of who is in which group. The respondent who flips "Tails" and answers truthfully knows that he or she cannot be distinguished from someone who got "Heads" in the coin toss. Hopefully, this is enough to encourage respondents to answer truthfully. As we will learn later in the course, the surveyor can then use probability methods to estimate the proportion of respondents who admit they used illegal drugs in this scenario, while being unable to identify exactly which respondents have been drug abusers.

Besides using the randomized response method, surveyors may encourage honest answers from respondents in various other ways. Tactful wording of questions can be very helpful. Giving people a feeling of anonymity by having them complete questionnaires via computer, rather than paper and pencil, is another commonly used technique.

Did I Get This?: Sample Surveys

Let's summarize

- A sample survey is a type of observational study in which respondents assess variables' values (often by giving an opinion).
- Open questions are less restrictive, but responses are more difficult to summarize.
- Closed questions may be biased by the options provided.
- Closed questions should permit options such as "other:_____" and/or "not sure" if those options may apply.
- Questions should be worded neutrally.
- Earlier questions should not deliberately influence responses to later questions.
- Questions shouldn't be confusing or complicated.
- Survey method and questions should be carefully designed to elicit honest responses if there are sensitive issues involved.

Sampling

CO-2: Differentiate among different sampling methods and discuss their strengths and limitations.

Video

Video: Sampling (12:38)

Sampling Plans

As mentioned in the introduction to this unit, we will begin with the first stage of data production — sampling. Our discussion will be framed around the following examples:

Suppose you want to determine the musical preferences of all students at your university, based on a sample of students. Here are some examples of the many possible ways to pursue this problem.

EXAMPLES: Sampling

Example 1: Post a music-lovers' survey on a university Internet bulletin board, asking students to vote for their favorite type of music.

This is an example of a **volunteer sample**, where individuals have selected themselves to be included. Such a sample is almost guaranteed to be **biased**. In general, volunteer samples tend to be comprised of individuals who have a particularly strong opinion about an issue, and are looking for an opportunity to voice it. Whether the variable's values obtained from such a sample are over- or under-stated, and to what extent, cannot be determined. As a result, data obtained from a voluntary response sample is quite useless when you think about the "Big Picture," since the sampled individuals only provide information about themselves, and we **cannot generalize to any larger group at all.**

Comment:

- It should be mentioned that in some cases volunteer samples are the only ethical way to obtain a sample. In medical studies, for example, in which new treatments are tested, subjects must choose to participate by signing a consent form that highlights the potential risks and benefits. As we will discuss in the next topic on study design, a volunteer sample is not so problematic in a study conducted for the purpose of comparing several treatments.

Example 2: Stand outside the Student Union, across from the Fine Arts Building, and ask the students passing by to respond to your question about musical preference.

This is an example of a **convenience sample**, where individuals happen to be at the right time and place to suit the schedule of the researcher. Depending on what variable is being studied, it may be that a convenience sample provides a fairly representative group. However, there are often subtle reasons why the sample's results are **biased**. In this case, the proximity to the Fine Arts Building might result in a disproportionate number of students favoring classical music. A convenience sample may also be susceptible to bias because certain types of individuals are more likely to be selected than others. In the extreme, some convenience samples are designed in such a way that certain individuals have no chance at all of being selected, as in the next example.

Example 3: Ask your professors for email rosters of all the students in your classes. Randomly sample some addresses, and email those students with your question about musical preference.

Here is a case where the **sampling frame** — list of potential individuals to be sampled — does not match the population of interest. The population of interest consists of all students at the university, whereas the sampling frame consists of only your classmates. There may be **bias** arising because of this discrepancy. For example, students with similar majors will tend to take the same classes as you, and their musical preferences may also be somewhat different from those of the general population of students. It is always best to have the sampling frame match the population as closely as possible.

Example 4: Obtain a student directory with email addresses of all the university's students, and send the music poll to every 50th name on the list.

This is called **systematic sampling**. It may not be subject to any clear bias, but it would not be as safe as taking a random sample.

If individuals are sampled completely at random, and without replacement, then each group of a given size is just as likely to be selected as all the other groups of that size. This is called a **simple random sample (SRS)**. In contrast, a systematic sample would not allow for sibling students to be selected, because of having the same last name. In a simple random sample, sibling students would have just as much of a chance of both being selected as any other pair of students. Therefore, there may be subtle sources of bias in using a systematic sampling plan.

Example 5: Obtain a student directory with email addresses of **all** the university's students, and send your music poll to a **simple random sample** of students.

As long as all of the students respond, then the sample is **not subject to any bias**, and should succeed in being representative of the population of interest.

But what if only 40% of those selected email you back with their vote?

The results of this poll would not necessarily be representative of the population, because of the potential problems associated with **volunteer response**. Since individuals are not compelled to respond, often a relatively small subset take the trouble to participate. Volunteer response is not as problematic as a volunteer sample (presented in example 1 above), but there is still a danger that those who do respond are different from those who don't, with respect to the variable of interest. An improvement would be to follow up with a second email, asking politely for the students' cooperation. This may boost the response rate, resulting in a sample that is fairly representative of the entire population of interest, and it may be the best that you can do, under the circumstances. **Nonresponse** is still an issue, but at least you have managed to reduce its impact on your results.

So far we've discussed several sampling plans, and determined that a simple random sample is the only one we discussed that is not subject to any bias.

A simple random sample is the easiest way to base a selection on randomness. There are other, more sophisticated, sampling techniques that utilize randomness that are often preferable in real-life circumstances. Any plan that relies on random selection is called a **probability sampling plan (or technique)**. The following three probability sampling plans are among the most commonly used:

- **Simple Random Sampling** is, as the name suggests, the simplest probability sampling plan. It is equivalent to "selecting names out of a hat." Each individual has the same chance of being selected.
- **Cluster Sampling** — This sampling technique is used when our population is naturally divided into groups (which we call clusters). For example, all the students in a university are divided into majors; all the nurses in a certain city are divided into hospitals; all registered voters are divided into precincts (election districts). In cluster sampling, we take a random sample of clusters, and use all the individuals within the selected clusters as our sample. For example, in order to get a sample of high-school seniors from a certain city, you choose 3 high schools at random from among all the high schools in that city, and use all the high school seniors in the three selected high schools as your sample.
- **Stratified Sampling** — Stratified sampling is used when our population is naturally divided into sub-populations, which we call stratum (plural: strata). For example, all the students in a certain college are divided by gender or by year in college; all the registered voters in a certain city are divided by race. In stratified sampling, we choose a simple random sample from each stratum, and our sample consists of all these simple random samples put together. For example, in order to get a random sample of high-school seniors from a certain city, we choose a random sample of 25 seniors from each of the high schools in that city. Our sample consists of all these samples put together.

Each of those probability sampling plans, if applied correctly, are not subject to any bias, and thus produce samples that represent well the population from which they were drawn.

Comment: Cluster vs. Stratified

- Students sometimes get confused about the difference between cluster sampling and stratified sampling. Even though both methods start out with the population somehow divided into groups, the two methods are very different.

- In cluster sampling, we take a random sample of whole groups of individuals taking everyone in that group but not all groups are taken), while in stratified sampling we take a simple random sample from each group (and all groups are represented).
- For example, say we want to conduct a study on the sleeping habits of undergraduate students at a certain university, and need to obtain a sample. The students are naturally divided by majors, and let's say that in this university there are 40 different majors.
  - In cluster sampling, we would randomly choose, say, 5 majors (groups) out of the 40, and use all the students in these five majors as our sample.
  - In stratified sampling, we would obtain a random sample of, say, 10 students from each of the 40 majors (groups), and use the 400 chosen students as the sample.
  - Clearly in this example, stratified sampling is much better, since the major of the student might have an effect on the student's sleeping habits, and so we would like to make sure that we have representatives from all the different majors. We'll stress this point again following the example and activity.

✔ EXAMPLE:

Suppose you would like to study the job satisfaction of hospital nurses in a certain city based on a sample. Besides taking a simple random sample, here are two additional ways to obtain such a sample.

1. Suppose that the city has 10 hospitals. Choose one of the 10 hospitals at random and interview all the nurses in that hospital regarding their job satisfaction. This is an example of cluster sampling, in which the hospitals are the clusters.

2. Choose a random sample of 50 nurses from each of the 10 hospitals and interview these 50 * 10 = 500 regarding their job satisfaction. This is an example of stratified sampling, in which each hospital is a stratum.

Cluster or Stratified — which one is better?

Let's go back and revisit the job satisfaction of hospital nurses example and discuss the pros and cons of the two sampling plans that are presented. Certainly, it will be much easier to conduct the study using the cluster sample, since all interviews are conducted in one hospital as opposed to the stratified sample, in which the interviews need to be conducted in 10 different hospitals. However, the hospital that a nurse works in probably has a direct impact on his/her job satisfaction, and in that sense, getting data from just one hospital might provide biased results. In this case, it will be very important to have representation from all the city hospitals, and therefore the stratified sample is definitely preferable. On the other hand, say that instead of job satisfaction, our study focuses on the age or weight of hospital nurses.

In this case, it is probably not as crucial to get representation from the different hospitals, and therefore the more easily obtained cluster sample might be preferable.

Comment:

- Another commonly used sampling technique is **multistage sampling**, which is essentially a "complex form" of cluster sampling. When conducting cluster sampling, it might be unrealistic, or too expensive to sample **all** the individuals in the chosen clusters. In cases like this, it would make sense to have another stage of sampling, in which you choose a sample from each of the randomly selected clusters, hence the term multistage sampling.

For example, say you would like to study the exercise habits of college students in the state of California. You might choose 8 colleges (clusters) at random, but you are certainly not going to use all the students in these 8 colleges as your sample. It is simply not realistic to conduct your study that way. Instead you move on to stage 2 of your sampling plan, in which you choose a random sample of 100 males and a random sample of 100 females from each of the 8 colleges you selected in stage 1.

So in total you have 8 * (100+100) = 1,600 college students in your sample.

In this case, stage 1 was a cluster sample of 8 colleges and stage 2 was a stratified sample within each college where the stratum was gender.

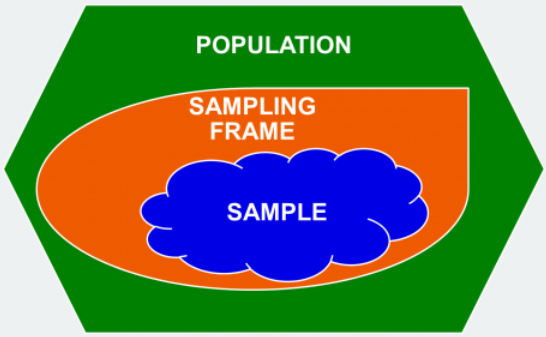
Multistage sampling can have more than 2 stages. For example, to obtain a random sample of physicians in the United States, you choose 10 states at random (stage 1, cluster). From each state you choose at random 8 hospitals (stage 2, cluster). Finally, from each hospital, you choose 5 physicians from each sub-specialty (stage 3, stratified).

Did I Get This?: Sampling

Overview So Far

We have defined the following:

Sampling Frame: List of potential individuals to be sampled. We want the sampling frame to match the population as closely as possible. The sampling frame is embedded within the population and the sample is embedded inside the sampling frame.



Biased Sample: A sample that produces data that is not representative because of the systematic under- or over-estimation of the values of the variable of interest.

Volunteer Sample: Individuals have selected themselves to be included.

Convenience Sample: Individuals happen to be at the right time and place to suit the schedule of the researcher

Systematic Sample: Starting from a randomly chosen individual in the ordered sampling frame, select every i-th individual to be included in the sample.

Simple Random Sample (SRS): Individuals are sampled completely at random, and without replacement. The result is that EVERY group of a given size is **just as likely to be selected** as all the other groups of that size. Each individual is also equally likely to be chosen.

Cluster Sampling: Used when "natural" groupings are evident in a statistical population and each group is generally representative of the population. In this technique, the total population is divided into these groups (or **clusters**) and a **sample of these groups** is selected. For example randomly selecting courses from all courses and surveying ALL students in selected courses.

Stratified Sampling: When subpopulations within an overall population vary, it can be advantageous to **take samples from each subpopulation (stratum) independently.** For example, take a random sample of males and a separate random sample of females.

Nonresponse: Individuals selected to participate do not respond or refuse to participate.

Sample Size

So far, we have made no mention of sample size. Our first priority is to make sure the sample is representative of the population, by using some form of probability sampling plan. Next, we must keep in mind that in order to get a more precise idea of what values are taken by the variable of interest for the entire population, a larger sample does a better job than a smaller one. We will discuss the issue of sample size in more detail in the Inference unit, and we will actually see how changes in the sample size affect the conclusions we can draw about the population.

EXAMPLE:

Suppose hospital administrators would like to find out how the staff would rate the quality of food in the hospital cafeteria. Which of the four sampling plans below would be best?

1. The person responsible for polling stands outside the cafeteria door and asks the next 5 staff members who come out to give the food a rating on a scale of 1 to 10.

2. The person responsible for polling stands outside the cafeteria door and asks the next 50 staff members who come out to give the food a rating on a scale of 1 to 10.

3. The person responsible for polling takes a random sample of 5 staff members from the list of all those employed at the hospital and asks them to rate the cafeteria food on a scale of 1 to 10.

4. The person responsible for polling takes a random sample of 50 staff members from the list of all those employed at the hospital and asks them to rate the cafeteria food on a scale of 1 to 10.

Plans 1 and 2 would be biased in favor of higher ratings, since staff members with unfavorable opinions about cafeteria food would be likely to eat elsewhere. Plan 3, since it is random, would be unbiased. However, with such a small sample, you run the risk of including people who provide unusually low or unusually high ratings. In other words, the average rating could vary quite a bit depending on who happens to be included in that small sample. Plan 4 would be best, as the participants have been chosen at random to avoid bias and the larger sample size provides more information about the opinions of all hospital staff members.

EXAMPLE:

Suppose a student enrolled in a statistics course is required to complete and turn in several hundred homework problems throughout the semester. The teaching assistant responsible for grading suggests the following plan to the course professor: instead of grading all of the problems for each student, he will grade a random sample of problems.

His first offer, to grade a random sample of just 3 problems for each student, is not well-received by the professor, who fears that such a small sample may not provide a very precise estimate of a student's overall homework performance.

Students are particularly concerned that the random selection may happen to include one or two problems on which they performed poorly, thereby lowering their grade.

The next offer, to grade a random sample of 25 problems for each student, is deemed acceptable by both the professor and the students.

Comment:

- In practice, we are confronted with many trade-offs in statistics. A larger sample is more informative about the population, but it is also more costly in terms of time and money. Researchers must make an effort to keep their costs down, but still obtain a sample that is large enough to allow them to report fairly precise results.

Learn By Doing: Sampling (Software)

Let's Summarize

Our goal, in statistics, is to use information from a sample to draw conclusions about the larger group, called the population. The **first step** in this process is to **obtain a sample** of individuals that are truly representative of the population. If this step is not

carried out properly, then the sample is subject to bias, a systematic tendency to misrepresent the variables of interest in the population.

Bias is almost guaranteed if a **volunteer sample** is used. If the individuals select themselves for the study, they are often different in an important way from the individuals who did not volunteer.

A **convenience sample**, chosen because individuals were in the right place at the right time to suit the researcher, may be different from the general population in a subtle but important way. However, for certain variables of interest, a convenience sample may still be fairly representative.

The **sampling frame** of individuals from whom the sample is actually selected should match the population of interest; bias may result if parts of the population are systematically excluded.

Systematic sampling takes an organized (but not random) approach to the selection process, as in picking every 50th name on a list, or the first product to come off the production line each hour. Just as with convenience sampling, there may be subtle sources of bias in such a plan, or it may be adequate for the purpose at hand.

Most studies are subject to some degree of **nonresponse**, referring to individuals who do not go along with the researchers' intention to include them in a study. If there are too many non-respondents, and they are different from respondents in an important way, then the sample turns out to be biased.

In general, bias may be eliminated (in theory), or at least reduced (in practice), if researchers do their best to implement a **probability sampling plan** that utilizes **randomness**.

The most basic probability sampling plan is a **simple random sample**, where every group of individuals has the same chance of being selected as every other group of the same size. This is achieved by sampling at random and without replacement.

In a **cluster sample**, groups of individuals are randomly selected, such as all people in the same household. In a cluster sample, all members of each selected group participate in the study.

A **stratified sample** divides the population into groups called strata before selecting study participants at random from within those groups.

Multistage sampling makes the sampling process more manageable by working down from a large population to successively smaller groups within the population, taking advantage of stratifying along the way, and sometimes finishing up with a cluster sample or a simple random sample.

Assuming the various sources of bias have been avoided, researchers can learn more about the variables of interest for the population by taking **larger samples**. The "extreme" (meaning, the largest possible sample) would be to study every single individual in the population (the goal of a census), but in practice, such a design is rarely feasible. Instead, researchers must try to obtain the largest sample that fits in their budget (in terms of both time and money), and must take great care that the sample is truly representative of the population of interest.

We will further discuss the topic of sample size when we cover sampling distributions and inferential statistics.

In this short section on sampling, we learned various techniques by which one can choose a sample of individuals from an entire population to collect data from. This is seemingly a simple step in the big picture of statistics, but it turns out that it has a crucial effect on the conclusions we can draw from the sample about the entire population (i.e., inference).

Caution

Generally speaking, a probability sampling plan (such as a simple random sample, cluster, or stratified sampling) will result in a nonbiased sample, which can be safely used to make inferences. Moreover, the inferential procedures that we will learn later in this course assume that the sample was chosen at random.

That being said, other (nonrandom) sampling techniques are available, and sometimes using them is the best we can do. It is important, though, when these techniques are used, to be aware of the types of bias that they introduce, and thus the limitations of the conclusions that can be drawn from the resulting samples.

Summary (Unit 2)

In this unit, we discussed the first step in the big picture of statistics — **production of data.**

Production of data happens in two stages: **sampling** and **study design**.

Our goal in sampling is to get a **sample that represents the population of interest well**, so that when we get to the inference stage, making conclusions based on this sample about the entire population will make sense.

We discussed several biased sampling plans, but also introduced the "family" of probability sampling plans, the simplest of which is the **simple random sample**, that (at least in theory) are supposed to provide a sample that is not subject to any biases.

In the section on study design, we introduced 3 types of design: observational study, controlled experiment, and sample survey.

We distinguished among different types of studies and learned the details of each type of study design. By doing so, we also expanded our understanding of the issue of establishing causation that was first discussed in the previous unit of the course. In the Exploratory Data Analysis unit, we learned that in general, association does not imply causation, due to the fact that lurking variables might be responsible for the association we observe, which means we cannot establish that there is a causal relationship between our "explanatory" variable and our response variable.

In this unit, we completed the causation puzzle by learning under what circumstances an observed association between variables CAN be interpreted as causation.

We saw that in observational studies, the best we can do is to control for what we think might be potential lurking variables, but we can never be sure that there aren't any others that we didn't anticipate. Therefore, we can come closer to establishing causation, but never really establish it.

The only way we can, at least in theory, eliminate the effect of (or control for) ALL lurking variables is by conducting a randomized controlled experiment, in which subjects are randomly assigned to one of the treatment groups. Only in this case can we interpret an observed association as causation.

Obviously, due to ethical or other practical reasons, not every study can be conducted as a randomized experiment. Where possible, however, a double-blind randomized controlled experiment is about the best study design we can use.

Another very common study design is the survey. While a survey is a special kind of observational study, it really is treated as a separate design, since it is so common and is the type of study that the general public is most often exposed to (polls). It is important that we be aware of the fact that the wording, ordering, or type of questions asked in a poll could have an impact on the response. In order for a survey's results to be reliable, these issues should be carefully considered when the survey is designed.

We saw that with **observational studies** it is **difficult to establish** convincing evidence of a **causal relationship**, because of lack of control over outside variables (called lurking variables). Other pitfalls that may arise are that individuals' behaviors may be affected if they know they are participating in an observational study, and that individuals' memories may be faulty if they are asked to recall information from the past.

Experiments allow researchers to take control of lurking variables by **randomized assignment to treatments**, which helps provide more convincing evidence of causation. The design may be enhanced by making sure that subjects and/or researchers are **blind** to who receives what treatment. Depending on what relationship is being researched, it may be difficult to design an experiment whose setting is realistic enough that we can safely generalize the conclusions to real life.

Another reason that observational studies are utilized rather than experiments is that certain explanatory variables — such as income or alcohol intake — either cannot or should not be controlled by researchers.

Sample surveys are occasionally used to examine relationships, but often they assess values of many separate variables, such as respondents' **opinions** on various matters. Survey questions should be designed carefully, in order to ensure unbiased assessment of the variables' values.

Throughout this unit, we established guidelines for the ideal production of data, which should be held as standards to strive for. Realistically, however, it is rarely possible to carry out a study which is completely free of flaws. Therefore, common sense must frequently be applied in order to decide which imperfections we can live with, and which ones could completely undermine a study's results.

(Optional) Outside Reading: Little Handbook – Design & Sampling (one long & one short)

CHAPTER OVERVIEW

Unit 3A: Probability

CO-1: Describe the roles biostatistics serves in the discipline of public health.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

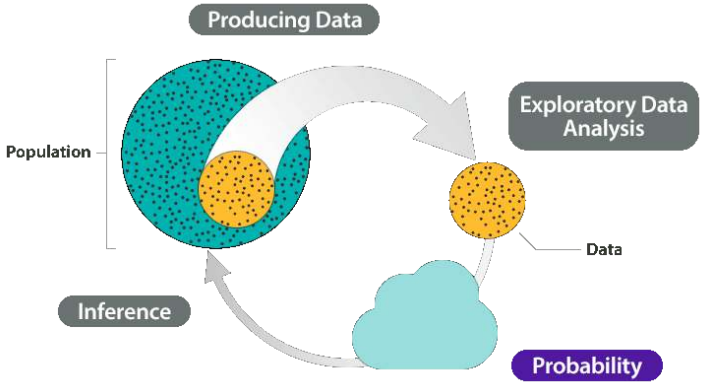
Video: Unit 3A: Introduction (5:36)

Review of the Big Picture

Learning Objectives

LO 1.3: Identify and differentiate between the components of the Big Picture of Statistics

Recall the Big Picture — the four-step process that encompasses statistics (as it is presented in this course):



So far, we've discussed the first two steps:

Producing data — how data are obtained, and what considerations affect the data production process.

Exploratory data analysis — tools that help us get a first feel for the data, by exposing their features using visual displays and numerical summaries which help us explore distributions, compare distributions, and investigate relationships.

(Recall that the structure of this course is such that Exploratory Data Analysis was covered first, followed by Producing Data.)

Our eventual goal is **Inference** — drawing reliable conclusions about the population based on what we've discovered in our sample.

In order to really understand how inference works, though, we first need to talk about **Probability**, because it is the underlying foundation for the methods of statistical inference.

The probability unit starts with an introduction, which will give you some motivating examples and an intuitive and informal perspective on probability.

Why do we need to understand probability?

- We often want to estimate the chance that an event (of interest to us) will occur.

- Many values of interest are probabilities or are derived from probabilities, for example, prevalence rates, incidence rates, and sensitivity/specificity of tests for disease.
- Plus!! Inferential statistics relies on probability to
  - Test hypotheses
  - Estimate population values, such as the population mean or population proportion.

Probability and Inference

We will use an example to try to explain why probability is so essential to inference.

First, here is the **general idea:**

As we all know, the way statistics works is that we use a sample to learn about the population from which it was drawn. Ideally, the sample should be random so that it represents the population well.

Recall from the discussion about sampling that **when we say that a random sample represents the population well we mean that there is no inherent bias** in this sampling technique.

It is important to acknowledge, though, that this does not mean that all random samples are necessarily "perfect." Random samples are still random, and therefore no random sample will be exactly the same as another.

One random sample may give a fairly accurate representation of the population, while another random sample might be "off," purely due to chance.

Unfortunately, when looking at a particular sample (which is what happens in practice), we will never know how much it differs from the population.

This **uncertainty** is where **probability** comes into the picture. This gives us a way to draw conclusions about the population in the face of the uncertainty that is generated by the use of a random sample.

> **We use probability to quantify how much we expect random samples to vary**.

The following example will illustrate this important point.

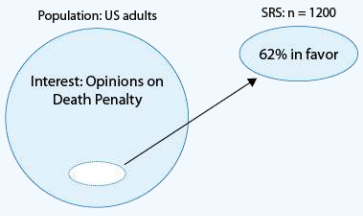
✔ EXAMPLE:

Suppose that we are interested in estimating the percentage of U.S. adults who favor the death penalty.

In order to do so, we choose a random sample of 1,200 U.S. adults and ask their opinion: either in favor of or against the death penalty.

We find that 744 out of the 1,200, or 62%, are in favor. (Comment: although this is only an example, this figure of 62% is quite realistic, given some recent polls).

Here is a picture that illustrates what we have done and found in our example:



Our goal here is inference — to learn and draw conclusions about the opinions of the entire population of U.S. adults regarding the death penalty, based on the opinions of only 1,200 of them.

Can we conclude that 62% of the population favors the death penalty?

- Another random sample could give a very different result. So we are uncertain.

But since our sample is random, we know that our uncertainty is due to chance, and not due to problems with how the sample was collected.

So we can use probability to describe the likelihood that our sample is within a desired level of precision.

For example, probability can answer the question, "How likely is it that our sample estimate is no more than 3% from the true percentage of all U.S. adults who are in favor of the death penalty?"

The answer to this question (which we find using probability) is obviously going to have an important impact on the confidence we can attach to the inference step.

In particular, if we find it quite unlikely that the sample percentage will be very different from the population percentage, then we have a lot of confidence that we can draw conclusions about the population based on the sample.

In the health sciences, a comparable situation to the death penalty example would be when we wish to determine the **prevalence** of a certain disease or condition.

> In epidemiology, the **prevalence** of a health-related state (typically disease, but also other things like smoking or seat belt use) in a statistical population is defined as the total number of cases in the population, divided by the number of individuals in the population.

As we will see, this is a form of probability.

In practice, we will need to estimate the prevalence using a sample and in order to make inferences about the population from a sample, we will need to understand probability.

✔ EXAMPLE:

The CDC estimates that in 2011, 8.3% of the U.S. population have diabetes. In other words, the CDC estimates the prevalence of diabetes to be 8.3% in the U.S.

> Fact Sheet on Diabetes from the CDC.

There are numerous statistics and graphs reported in this document you should now understand!!

Other common probabilities used in the health sciences are

- (Cumulative) **Incidence**: the probability that a person with no prior disease will develop disease over some specified time period
- **Sensitivity** of a diagnostic or screening test: the probability the person tests positive, given the person has the disease. **Specificity** of a diagnostic or screening test: the probability the person tests negative, given the person does not have the disease. As well as **predictive value positive**, **predictive value negative**, **false positive rate**, **false negative rate**.
- **Survival probability**: the probability an individual survives beyond a certain time

Basic Probability Rules
Conditional Probability and Independence
Introduction to Probability
Summary (Unit 3)

Basic Probability Rules

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.4: Relate the probability of an event to the likelihood of this event occurring.

Learning Objectives

LO 6.5: Apply the relative frequency approach to estimate the probability of an event.

Learning Objectives

LO 6.6: Apply basic logic and probability rules in order to find the empirical probability of an event.

Video

Video: Basic Probability Rules (25:17)

In the previous section, we introduced **probability** as a way to **quantify the uncertainty** that arises from conducting experiments using a random sample from the population of interest.

We saw that the **probability** of an event (for example, the event that a randomly chosen person has blood type O) **can be estimated** by the **relative frequency** with which the event occurs **in a long series of trials**. So we would collect data from lots of individuals to estimate the probability of someone having blood type O.

In this section, we will **establish** the **basic methods and principles for finding probabilities** of events.

We will also cover some of the **basic rules of probability** which can be used to **calculate probabilities**.

Introduction

We will begin with a classical probability example of tossing a fair coin three times.

Since heads and tails are equally likely for each toss in this scenario, each of the possibilities which can result from three tosses will also be equally likely so that we can list all possible values and use this list to calculate probabilities.

Since our focus in this course is on data and statistics (not theoretical probability), in most of our future problems we will use a summarized dataset, usually a frequency table or two-way table, to calculate probabilities.

EXAMPLE: Toss a fair coin three times

Let's list each possible outcome (or possible result):

{HHH, THH, HTH, HHT, HTT, THT, TTH, TTT}

Now let's define the following events:

Event A: "Getting no H"

Event B: "Getting exactly one H"

Event C: "Getting at least one H"

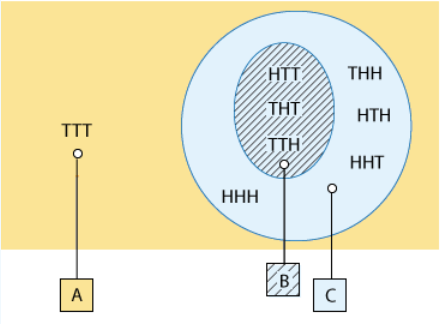
Note that each event is indeed a statement about the outcome that the experiment is going to produce. In practice, each event corresponds to some collection (subset) of the possible outcomes.

Event A: "Getting no H" → TTT

Event B: "Getting exactly one H" → HTT, THT, TTH

Event C: "Getting at least one H" → HTT, THT, TTH, THH, HTH, HHT, HHH

Here is a visual representation of events A, B and C.



From this visual representation of the events, it is easy to see that event B is totally included in event C, in the sense that every outcome in event B is also an outcome in event C. Also, note that event A stands apart from events B and C, in the sense that they have no outcome in common, or no overlap. At this point these are only noteworthy observations, but as you'll discover later, they are very important ones.

What if we added the new event:

Event D: "Getting a T on the first toss" → THH, THT, TTH, TTT

How would it look if we added event D to the diagram above? (Link to the answer)

Remember, since H and T are equally likely on each toss, and since there are 8 possible outcomes, the probability of each outcome is 1/8.

See if you can answer the following questions using the diagrams and/or the list of outcomes for each event along with what you have learned so far about probability.

> **Learn By Doing:** Tossing a Fair Coin Three Times

If you were able to answer those questions correctly, you likely have a good instinct for calculating probability! Read on to learn how we will apply this knowledge.

If not, we will try to help you develop this skill in this section.

Comment:

- Note that in event C, "Getting at least one head" there is only one possible outcome which is missing, "Getting NO heads" = TTT. We will address this again when we talk about probability rules, in particular the complement rule. At this point, we just want you to think about how these two events are "opposites" in this scenario.

It is VERY important to realize that just because we can list out the possible outcomes, this does not imply that each outcome is equally likely.

This is the (funny) message in the Daily Show clip we provided on the previous page. But let's think about this again. In that clip, Walter is claiming that since there are two possible outcomes, the probability is 0.5. The two possible outcomes are

- The world will be destroyed due to use of the large hadron collider
- The world will NOT be destroyed due to use of the large hadron collider

Hopefully it is clear that these two outcomes are not equally likely!!

Let's consider a more common example.

✔ EXAMPLE: Birth Defects

Suppose we randomly select three children and we are interested in the probability that none of the children have any birth defects.

We use the notation D to represent a child was born with a birth defect and N to represent the child born with NO birth defect. We can list the possible outcomes just as we did for the coin toss, they are:

{DDD, NDD, DND, DDN, DNN, NDN, NND, NNN}

Are the events DDD (all three children are born with birth defects) and NNN (none of the children are born with birth defects) equally likely?

It should be reasonable to you that **P(NNN) is much larger than P(DDD)**.

This is because P(N) and P(D) are not equally likely events.

It is rare (certainly not 50%) for a randomly selected child to be born with a birth defect.

Rules of Probability

Now we move on to learning some of the basic rules of probability.

Fortunately, these rules are very intuitive, and as long as they are applied systematically, they will let us solve more complicated problems; in particular, those problems for which our intuition might be inadequate.

Since **most of the probabilities you will be asked to find can be calculated** using both

- **logic and counting**

and

- **the rules we will be learning,**

we give the following advice as a principle.

PRINCIPLE:

If you can calculate a probability using logic and counting you do not NEED a probability rule (although the correct rule can always be applied)

Probability Rule One

Our first rule simply reminds us of the basic property of probability that we've already learned.

The probability of an event, which informs us of the likelihood of it occurring, can range anywhere from 0 (indicating that the event will never occur) to 1 (indicating that the event is certain).

Probability Rule One:

- **For any event A, $0 \leq P(A) \leq 1$.**

NOTE: One practical use of this rule is that it can be used to identify any probability calculation that comes out to be more than 1 (or less than 0) as incorrect.

Before moving on to the other rules, let's first look at an example that will provide a context for illustrating the next several rules.

✔ EXAMPLE: Blood Types

As previously discussed, all human blood can be typed as O, A, B or AB.

In addition, the frequency of the occurrence of these blood types varies by ethnic and racial groups.

According to Stanford University's Blood Center (bloodcenter.stanford.edu), these are the probabilities of human blood types in the United States (the probability for type A has been omitted on purpose):

Blood type	O	A	B	AB
Probability	0.44	?	0.10	0.04

Motivating question for rule 2: A person in the United States is chosen at random. What is the probability of the person having blood type A?

Answer Our intuition tells us that since the four blood types O, A, B, and AB exhaust all the possibilities, their probabilities together must sum to 1, which is the probability of a "certain" event (a person has one of these 4 blood types for certain).

Since the probabilities of O, B, and AB together sum to 0.44 + 0.1 + 0.04 = **0.58**, the probability of type A must be the remaining **0.42** (1 – 0.58 = 0.42):

Blood type	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

Probability Rule Two

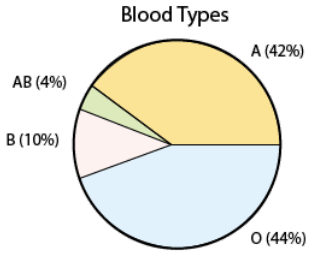
This example illustrates our second rule, which tells us that the probability of all possible outcomes together must be 1.

Probability Rule Two:

The sum of the probabilities of all possible outcomes is 1.

This is a good place to compare and contrast what we're doing here with what we learned in the Exploratory Data Analysis (EDA) section.

- Notice that in this problem we are essentially focusing on a single categorical variable: blood type.
- We summarized this variable above, as we summarized single categorical variables in the EDA section, by listing what values the variable takes and how often it takes them.
- In EDA we used percentages, and here we're using probabilities, but the two convey the same information.
- In the EDA section, we learned that a pie chart provides an appropriate display when a single categorical variable is involved, and similarly we can use it here (using percentages instead of probabilities):



Even though what we're doing here is indeed similar to what we've done in the EDA section, there is a subtle but important difference between the underlying situations

- In EDA, we summarized data that were obtained from a **sample**of individuals for whom values of the variable of interest were recorded.
- Here, when we present the probability of each blood type, we have in mind the entire **population**of people in the United States, for which we are presuming to know the overall frequency of values taken by the variable of interest.

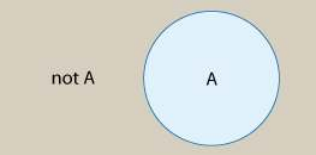
Did I Get This?: Probability Rule Two

Probability Rule Three

In probability and in its applications, we are frequently interested in finding out the probability that a certain event will **not** occur.

An important point to understand here is that "event A does not occur" is **a separate event** that consists of all the possible outcomes that are not in A and is called "**the complement event of A**."

Notation: we will write **"not A"** to denote the event that A does **not** occur. Here is a visual representation of how event A and its complement event "not A" together represent all possible outcomes.



Comment:

- Such a visual display is called a "Venn diagram." A Venn diagram is a simple way to visualize events and the relationships between them using rectangles and circles.

Rule 3 deals with the relationship between the probability of an event and the probability of its complement event.

Given that event A and event "not A" together make up all possible outcomes, and since rule 2 tells us that the sum of the probabilities of all possible outcomes is 1, the following rule should be quite intuitive:

Probability Rule Three (The Complement Rule):

- **P(not A) = 1 – P(A)**
- that is, the probability that an event does not occur is 1 minus the probability that it does occur.

EXAMPLE: Blood Types

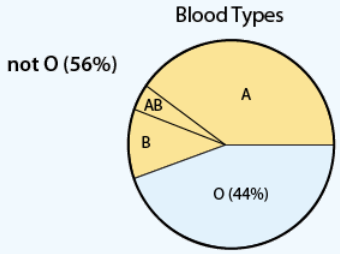
Back to the blood type example:

Blood type	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

Here is some additional information:

- A person with type **A** can donate blood to a person with type **A** or **AB**.
- A person with type **B** can donate blood to a person with type **B** or **AB**.
- A person with type **AB** can donate blood to a person with type **AB** only.
- A person with type **O** blood can donate to anyone.

What is the probability that a randomly chosen person cannot donate blood to everyone? In other words, what is the probability that a randomly chosen person does not have blood type O? We need to find P(not O). Using the Complement Rule, P(not O) = 1 – P(O) = 1 – 0.44 = 0.56. In other words, 56% of the U.S. population does not have blood type O:



Clearly, we could also find P(not O) directly by adding the probabilities of B, AB, and A.

Comment:

- Note that the Complement Rule, **P(not A) = 1 – P(A)** can be re-formulated as **P(A) = 1 – P(not A).**
  - **P(not A) = 1 – P(A)**
  - can be re-formulated as **P(A) = 1 – P(not A).**
  - This seemingly trivial algebraic manipulation has an important application, and actually captures the strength of the complement rule.
  - In some cases, when finding P(A) directly is very complicated, it might be much easier to find P(not A) and then just subtract it from 1 to get the desired P(A).
  - We will come back to this comment soon and provide additional examples.

> **Did I Get This?:** Probability Rule Three

Comments:

- The complement rule can be useful whenever it is easier to calculate the probability of the complement of the event rather than the event itself.
- Notice, we again used the phrase "at least one."
- Now we have seen that the complement of "at least one …" is "none … " or "no …." (as we mentioned previously in terms of the events being "opposites").
- In the above activity we see that
  - **P(NONE of these two side effects) = 1 – P(at least one of these two side effects)**
- This is a common application of the complement rule which you can often recognize by the phrase **"at least one"** in the problem.

Probabilities Involving Multiple Events

We will often be interested in finding probabilities involving multiple events such as

- **P(A or B) = P(event A occurs or event B occurs or both occur)**
- **P(A and B)= P(both event A occurs and event B occurs)**

A common issue with terminology relates to how we usually think of "or" in our daily life. For example, when a parent says to his or her child in a toy store "Do you want toy A or toy B?", this means that the child is going to get only one toy and he or she has to choose between them. Getting both toys is usually not an option.

In contrast:

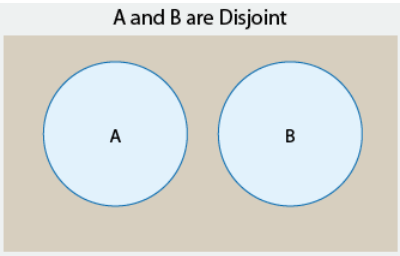
> **In probability, "OR" means either one or the other or both.**
>
> and so **P(A or B) = P(event A occurs or event B occurs or BOTH occur)**

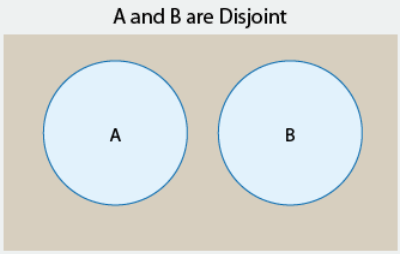
Having said that, it should be noted that there are some cases where it is simply impossible for the two events to both occur at the same time.

Probability Rule Four

The distinction between events that can happen together and those that cannot is an important one.

> **Disjoint:** Two events that cannot occur at the same time are called disjoint or mutually exclusive. (We will use disjoint.)





It should be clear from the picture that

- in the first case, where the events are **NOT disjoint, P(A and B) ≠ 0**
- in the second case, where the events **ARE disjoint, P(A and B) = 0.**

Here are two examples:

EXAMPLE:

Consider the following two events:

A — a randomly chosen person has blood type A, and

B — a randomly chosen person has blood type B.

In rare cases, it is possible for a person to have more than one type of blood flowing through his or her veins, but for our purposes, we are going to assume that each person can have only one blood type. Therefore, it is impossible for the events A and B to occur together.

- **Events A and B are DISJOINT**

On the other hand …

EXAMPLE:

Consider the following two events:

A — a randomly chosen person has blood type A

B — a randomly chosen person is a woman.

In this case, it **is possible** for events A and B to occur together.

- **Events A and B are NOT DISJOINT.**

The Venn diagrams suggest that another way to think about disjoint versus not disjoint events is that disjoint events **do not overlap**. They do not share any of the possible outcomes, and therefore cannot happen together.

On the other hand, events that are not disjoint are overlapping in the sense that they share some of the possible outcomes and therefore can occur at the same time.

We now begin with a simple rule for finding P(A or B) for disjoint events.

Probability Rule Four (The Addition Rule for Disjoint Events):

- **If A and B are disjoint events, then P(A or B) = P(A) + P(B).**

Comment:

- When dealing with probabilities, the word "or" will always be associated with the operation of addition; hence the name of this rule, "The Addition Rule."

EXAMPLE: Blood Types

Recall the blood type example:

Blood type	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

Here is some additional information

- A person with type **A**can donate blood to a person with type **A** or **AB**.
- A person with type **B**can donate blood to a person with type **B** or **AB**.
- A person with type **AB**can donate blood to a person with type **AB**
- A person with type **O**blood can donate to anyone.

What is the probability that a randomly chosen person is a potential donor for a person with blood type A?

From the information given, we know that being a potential donor for a person with blood type A means having blood type A or O.

We therefore need to find P(A or O). Since the events A and O are disjoint, we can use the addition rule for disjoint events to get:

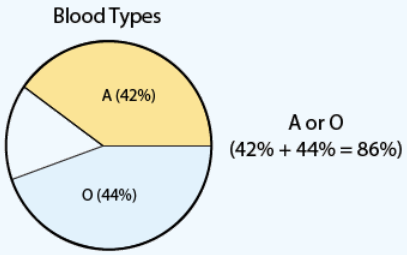
- **P(A or O) = P(A) + P(O) = 0.42 + 0.44 = 0.86.**

It is easy to see why adding the probability actually makes sense.

If 42% of the population has blood type A and 44% of the population has blood type O,

- then 42% + 44% = 86% of the population has either blood type A or O, and thus are potential donors to a person with blood type A.

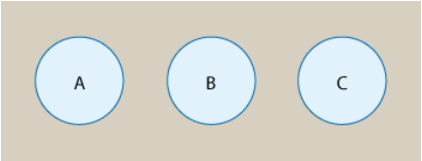
This reasoning about why the addition rule makes sense can be visualized using the pie chart below:



Learn By Doing: Probability Rule Four

Comment:

- The Addition Rule for Disjoint Events can naturally be extended to more than two disjoint events. Let's take three, for example. If A, B and C are three disjoint events



then P(A or B or C) = P(A) + P(B) + P(C). The rule is the same for any number of disjoint events.

Did I Get This?: Probability Rule Four

We are now finished with the first version of the Addition Rule (Rule four) which is the version restricted to disjoint events. Before covering the second version, we must first discuss P(A and B).

Finding P(A and B) using Logic

We now turn to calculating

- **P(A and B)= P(both event A occurs and event B occurs)**

Later, we will discuss the rules for calculating P(A and B).

First, we want to illustrate that a rule is not needed whenever you can determine the answer through logic and counting.

Special Case:

There is one special case for which we know what P(A and B) equals without applying any rule.

Learn by Doing: Finding P(A and B) #1

So, if events **A and B are disjoint**, then (by definition) **P(A and B)= 0.** But what if the events are not disjoint?

Recall that rule 4, the Addition Rule, has two versions. One is restricted to disjoint events, which we've already covered, and we'll deal with the more general version later in this module. The same will be true of probabilities involving AND

However, **except in special cases, we will rely on LOGIC to find P(A and B) in this course.**

Before covering any formal rules, let's look at an example where the events are not disjoint.

EXAMPLE: Periodontal Status and Gender

Learn by Doing: Periodontal Status and Gender

We like to ask probability questions similar to the previous example (using a two-way table based upon data) as this allows you to make connections between these topics and helps you keep some of what you have learned about data fresh in your mind.

Caution

Remember, our primary goal in this course is to analyze real-life data!

Probability Rule Five

We are now ready to move on to the extended version of the Addition Rule.

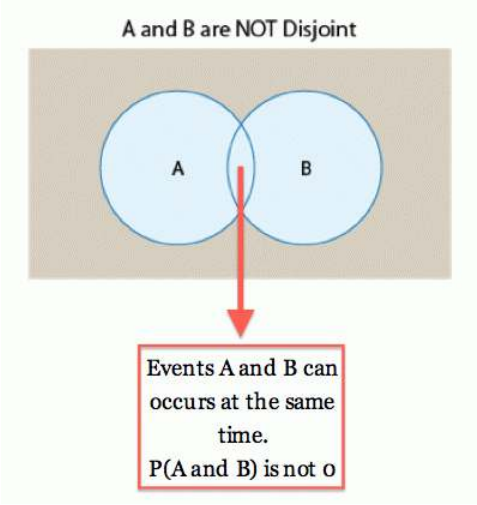
In this section, we will learn how to find P(A or B) when A and B are not necessarily disjoint.

- We'll call this extended version the "**General Addition Rule**" and state it as **Probability Rule Five**.

We will begin by stating the rule and providing an example similar to the types of problems we generally ask in this course. Then we will present a more another example where we do not have the raw data from a sample to work from.

As we witnessed in previous examples, when the two events are not disjoint, there is some overlap between the events.

- If we simply add the two probabilities together, we will get the wrong answer because we have counted some "probability" twice!
- Thus, we must subtract out this "extra" probability to arrive at the correct answer. The Venn diagram and the two-way tables are helpful in visualizing this idea.

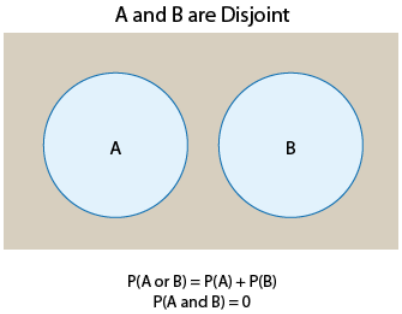


This rule is more general since it works for any pair of events (even disjoint events). Our advice is still to try to answer the question using logic and counting whenever possible, otherwise, we must be extremely careful to choose the correct rule for the problem.

PRINCIPLE:

If you can calculate a probability using logic and counting you do not NEED a probability rule (although the correct rule can always be applied)

Notice that, if A and B are disjoint, then P(A and B) = 0 and rule 5 reduces to rule 4 for this special case.



Let's revisit the last example:

EXAMPLE: Periodontal Status and Gender

Consider randomly selecting one individual from those represented in the following table regarding the periodontal status of individuals and their gender. Periodontal status refers to gum disease where individuals are classified as either healthy, have gingivitis, or have periodontal disease.

Count

		periodontal status			
		healthy	gingivitis	perio	Total
GENDER	male	1143	929	937	3009
	female	2607	1490	921	5018
Total		3750	2419	1858	8027

Let's review what we have learned so far. We can calculate any probability in this scenario if we can determine how many individuals satisfy the event or combination of events.

- P(Male) = 3009/8027 = 0.3749
- P(Female) = 5018/8027 = 0.6251
- P(Healthy) = 3750/8027 = 0.4672
- P(Not Healthy) = P(Gingivitis or Perio) = (2419 + 1858)/8027 = 4277/8027 = 0.5328
  We could also, calculate this using the complement rule: 1 – P(Healthy)

We also previously found that

- P(Male AND Healthy) = 1143/8027 = 0.1424

Count

		periodontal status			
		healthy	gingivitis	perio	Total
GENDER	male	1143	929	937	3009
	female	2607	1490	921	5018
Total		3750	2419	1858	8027

Recall rule 5, P(A or B) = P(A) + P(B) – P(A and B). We now use this rule to calculate P(Male OR Healthy)

- P(Male or Healthy) = P(Male) + P(Healthy) – P(Male and Healthy) = 0.3749 + 0.4672 – 0.1424 = 0.6997 or about 70%

We solved this question earlier by simply counting how many individuals are either Male or Healthy or both. The picture below illustrates the values we need to combine. We need to count

- All males
- All healthy individuals
- BUT, not count anyone twice!!

Count

		periodontal status			
		healthy	gingivitis	perio	Total
GENDER	male	1143	929	937	3009
	female	2607	1490	921	5018
Total		3750	2419	1858	8027

Using this logical approach we would find

- P(Male or Healthy) = (1143 + 929 + 937 + 2607)/8027 = 5616/8027 = 0.6996

We have a minor difference in our answers in the last decimal place due the rounding that occurred when we calculated P(Male), P(Healthy), and P(Male and Healthy) and then applied rule 5.

Clearly the answer is effectively the same, about 70%. If we carried our answers to more decimal places or if we used the original fractions, we could eliminate this small discrepancy entirely.

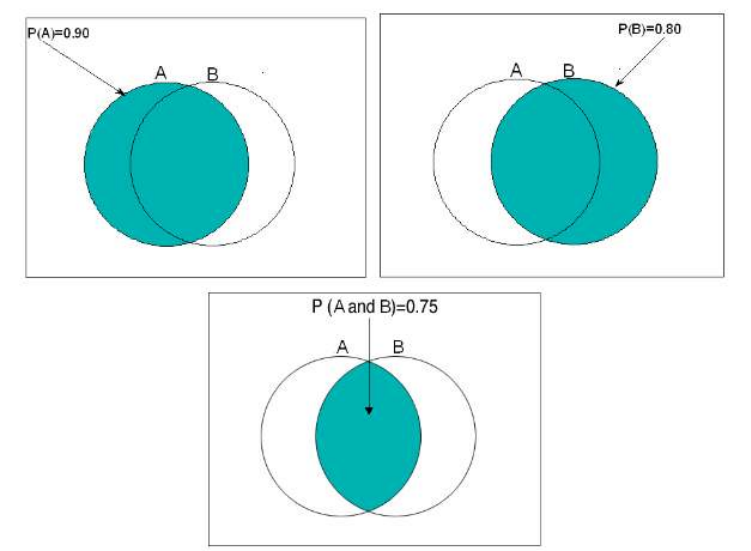
Let's look at one final example to illustrate Probability Rule 5 when the rule is needed – i.e. when we don't have actual data.

EXAMPLE: Important Delivery!

It is vital that a certain document reach its destination within one day. To maximize the chances of on-time delivery, two copies of the document are sent using two services, service A and service B. It is known that the probabilities of on-time delivery are:

- 0.90 for service A (**P(A) = 0.90**)
- 0.80 for service B (**P(B) = 0.80**)
- 0.75 for both services being on time (**P(A and B) = 0.75**)
  (Note that A and B are n**ot disjoint**. They can happen together with probability 0.75.)

The Venn diagrams below illustrate the probabilities P(A), P(B), and P(A and B) [not drawn to scale]:

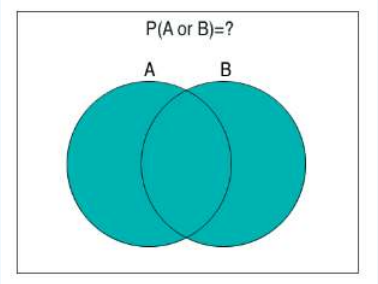


In the context of this problem, the obvious question of interest is:

- ***What is the probability of on-time delivery of the document using this strategy (of sending it via both services)?***

The document will reach its destination on time as long as it is delivered on time by service A or by service B or by both services. In other words, when event A occurs or event B occurs or both occur. so….

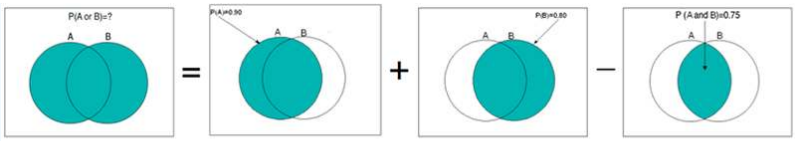
P(on time delivery using this strategy)= **P(A or B)**, which is represented the by the shaded region in the diagram below:



We can now

- use the three Venn diagrams representing **P(A), P(B) and P(A and B)**
- to see that we can find **P(A or B)** by adding **P(A)** (represented by the left circle) and **P(B)** (represented by the right circle),
- then subtracting **P(A and B) (represented by the overlap),** since we included it twice, once as part of P(A) and once as part of P(B).

This is shown in the following image:



We therefore get:

P(A or B) = P(A) + P(B) – P(A and B)

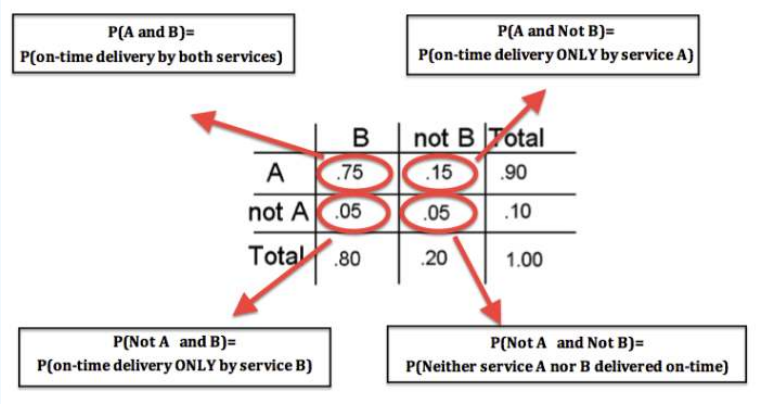
If we apply this to our example, we find that:

- **P(A or B)= P(on-time delivery using this strategy)= 0.90 + 0.80 – 0.75 = 0.95.**

So our strategy of using two delivery services increases our probability of on-time delivery to 0.95.

While the Venn diagrams were great to visualize the General Addition Rule, in cases like these it is much easier to display the information in and work with a two-way table of probabilities, much as we examined the relationship between two categorical variables in the Exploratory Data Analysis section.

We will simply show you the table, not how we derive it as you won't be asked to do this for us. You should be able to see that some logic and simple addition/subtraction is all we used to fill in the table below.



When using a two-way table, we must remember to look at the entire row or column to find overall probabilities involving only A or only B.

- P(A) = 0.90 means that in 90% of the cases when service A is used, it delivers the document on time. To find this we look at the total probability for the row containing A. In finding P(A), we do not know whether B happens or not.

	B	not B	Total
A	.75	.15	.90
not A	.05	.05	.10
Total	.80	.20	1.00

- P(B) = 0.80 means that in 80% of the cases when service B is used, it delivers the document on time. To find this we look at the total probability for the column containing B. In finding P(B), we do not know whether A happens or not.

	B	not B	Total
A	.75	.15	.90
not A	.05	.05	.10
Total	.80	.20	1.00

Comment

- When we used two-way tables in the Exploratory Data Analysis (EDA) section, it was to record values of two categorical variables for a concrete **sample** of individuals.
- In contrast, the information in a probability two-way table is for an entire **population**, and the values are rather abstract.
- If we had treated something like the delivery example in the EDA section, we would have recorded the actual numbers of on-time (and not-on-time) deliveries for **samples** of documents mailed with service A or B.
- In this section, the **long-term probabilities are presented as being known**.
- Presumably, the reported probabilities in this delivery example were based on relative frequencies recorded over many repetitions.

Interactive Applet: Probability Venn Diagram

Rounding Rule of Thumb for Probability:

Follow the following general guidelines in this course. If in doubt carry more decimal places. If we specify give exactly what is requested.

- **In general you should carry probabilities to at least 4 decimal places for intermediate steps.**
- **We often round our final answer to two or three decimal places.**
- **For extremely small probabilities, it is important to have 1 or two significant digits (non-zero digits), such as 0.000001 or 0.000034, etc.**

Many computer packages might display extremely small values using scientific notation such as

- 58×10^{-5} or 1.58 E^{-5} to represent 0.0000158

Let's Summarize

So far in our study of **probability**, you have been introduced to the sometimes counter-intuitive nature of probability and the fundamentals that underlie probability, such as a **relative frequency**.

We also gave you some tools to help you find the probabilities of events — namely the **probability rules**.

You probably noticed that the probability section was significantly different from the two previous sections; it has a much larger technical/mathematical component, so the results tend to be more of the "right or wrong" nature.

In the Exploratory Data Analysis section, for the most part, the computer took care of the technical aspect of things, and our tasks were to tell it to do the right thing and then interpret the results.

In probability, we do the work from beginning to end, from choosing the right tool (rule) to use, to using it correctly, to interpreting the results.

Here is a summary of the rules we have presented so far.

1. Probability Rule #1 states:

- **For any event A, $0 \le P(A) \le 1$**

2. Probability Rule #2 states:

- **The sum of the probabilities of all possible outcomes is 1**

3. The Complement Rule (#3) states that

- **P(not A) = 1 – P(A)**

or when rearranged

- **P(A) = 1 – P(not A)**

The latter representation of the Complement Rule is especially useful when we need to find probabilities of events of the sort "at least one of …"

4. The General Addition Rule (#5) states that for any two events,

- **P(A or B) = P(A) + P(B) – P(A and B),**

where, by P(A or B) we mean P(A occurs or B occurs or both).

In the special case of **disjoint** events, events that cannot occur together, the General Addition Rule can be reduced to the Addition Rule for Disjoint Events (#4), which is

- **P(A or B) = P(A) + P(B).** *

*ONLY use when you are CONVINCED the events are disjoint (they do NOT overlap)

5. The **restricted version** of the addition rule (for disjoint events) **can be easily extended** to more than two events.

6. So far, we have only found **P(A and B)** using logic and counting in simple examples

---

Conditional Probability and Independence

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.4: Relate the probability of an event to the likelihood of this event occurring.

Learning Objectives

LO 6.5: Apply the relative frequency approach to estimate the probability of an event.

Learning Objectives

LO 6.6: Apply basic logic and probability rules in order to find the empirical probability of an event.

Review: Unit 1 Case C-C

- In particular the idea of **conditional percentages** will be equivalent to the idea of **conditional probabilities** discussed in this section.

Video

Video: Conditional Probability and Independence (28:13)

In the last section, we established some of the basic rules of probability, which included:

- Basic Properties of Probability (Rule One and Rule Two)
- The Complement Rule (Rule Three)
- The Addition Rule for **Disjoint** Events (Rule Four)
- The General Addition Rule for which the events need not be disjoint (Rule Five)

In order to complete our set of rules, we still require two Multiplication Rules for finding P(A and B) and the important concepts of **independent events** and **conditional probability.**

We'll first introduce the idea of independent events, then introduce the Multiplication Rule for independent events which gives a way to find P(A **and** B) in cases when the events A and B are independent.

Next we will define conditional probability and use it to formalize our definition of independent events, which is initially presented only in an intuitive way.

We will then develop the General Multiplication Rule, a rule that will tell us how to find P(A **and** B) in cases when the events A and B are not necessarily independent.

We'll conclude with a discussion of probability applications in the health sciences.

Independent Events

Learning Objectives

LO 6.7: Determine whether two events are independent or dependent and justify your conclusion.

We begin with a verbal definition of independent events (later we will use probability notation to define this more precisely).

Independent Events:

- Two events A and B are said to be **independent** if the fact that one event has occurred **does not affect** the probability that the other event will occur.
- If whether or not one event occurs **does affect** the probability that the other event will occur, then the two events are said to be **dependent.**

Here are a few examples:

EXAMPLE:

A woman's pocket contains two quarters and two nickels.

She randomly extracts one of the coins and, after looking at it, replaces it before picking a second coin.

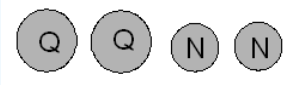
Let Q1 be the event that the first coin is a quarter and Q2 be the event that the second coin is a quarter.

Are Q1 and Q2 independent events?

- Why?

Since the first coin that was selected is **replaced**, whether or not Q1 occurred (i.e., whether the first coin was a quarter) has no effect on the probability that the second coin will be a quarter, P(Q2).

In either case (whether Q1 occurred or not), when she is selecting the second coin, she has in her pocket:



and therefore the **P(Q2) = 2/4 = 1/2** regardless of whether Q1 occurred.

EXAMPLE:

A woman's pocket contains two quarters and two nickels.

She randomly extracts one of the coins, and **without placing** it back into her pocket, she picks a second coin.

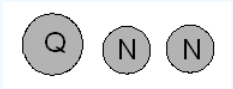
As before, let Q1 be the event that the first coin is a quarter, and Q2 be the event that the second coin is a quarter.

Are Q1 and Q2 independent events?

- Q1 and Q2 are **not independent**. They are **dependent.** Why?

Since the first coin that was selected is **not replaced,** whether Q1 occurred (i.e., whether the first coin was a quarter) **does affect** the probability that the second coin is a quarter, P(Q2).

If Q1 occurred (i.e., the first coin was a quarter), then when the woman is selecting the second coin, she has in her pocket:



- In this case, **P(Q2) = 1/3.**

However, **if Q1 has not occurred** (i.e., the first coin was not a quarter, but a nickel), then when the woman is selecting the second coin, she has in her pocket:

- In this case, **P(Q2) = 2/3.**

In these last two examples, we could actually have done some calculation in order to check whether or not the two events are independent or not.

Sometimes we can just use common sense to guide us as to whether two events are independent. Here is an example.

EXAMPLE:

Two people are selected simultaneously and at random from all people in the United States.

Let B1 be the event that one of the people has blue eyes and B2 be the event that the other person has blue eyes.

In this case, since they were chosen at random, whether one of them has blue eyes has no effect on the likelihood that the other one has blue eyes, and therefore **B1 and B2 are independent**.

On the other hand …

EXAMPLE:

A family has 4 children, two of whom are selected at random.

Let B1 be the event that one child has blue eyes, and B2 be the event that the other chosen child has blue eyes.

In this case, **B1 and B2 are not independent, since we know that eye color is hereditary.**

Thus, whether or not one child is blue-eyed will increase or decrease the chances that the other child has blue eyes, respectively.

Comments:

- It is quite common for students to initially get confused about the distinction between the idea of **disjoint events** and the idea of **independent events**. The purpose of this comment (and the activity that follows it) is to help students develop more understanding about these very different ideas.

The idea of **disjoint events** is about whether or not it is possible for the events to occur at the same time (see the examples on the page for Basic Probability Rules).

The idea of **independent events** is about whether or not the events affect each other in the sense that the occurrence of one event affects the probability of the occurrence of the other (see the examples above).

The following activity deals with the distinction between these concepts.

The purpose of this activity is to help you strengthen your understanding about the concepts of disjoint events and independent events, and the distinction between them.

Learn by Doing: Independent Events

Let's summarize the three parts of the activity:

- In Example 1: A and B are **not disjoint** and **independent**
- In Example 2: A and B are **not disjoint** and **not independent**
- In Example 3: A and B are **disjoint** and **not independent**.

Why did we leave out the case when the events are disjoint and independent?

The reason is that this case DOES NOT EXIST!

	A and B Independent	**A and B Not Independent**
A and B Disjoint	**DOES NOT EXIST**	Example 3
A and B Not Disjoint	Example 1	Example 2

If events are **disjoint then they must be not independent, i.e. they must be dependent events.**

Why is that?

- Recall: If A and B are disjoint then they cannot happen together.
- In other words, A and B being disjoint events implies that if event A occurs then B does not occur and vice versa.
- Well… if that's the case, knowing that event A has occurred dramatically changes the likelihood that event B occurs – that likelihood is zero.
- This implies that A and B are not independent.

Now that we understand the idea of independent events, we can finally get to rules for finding P(A and B) in the special case in which the events A and B are independent.

Later we will present a more general version for use when the events are not necessarily independent.

Multiplication Rule for Independent Events (Rule Six)

Learning Objectives

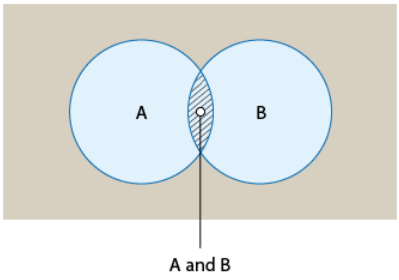
LO 6.8: Apply the multiplication rule for independent events to calculate P(A and B) for independent events.

We now turn to rules for calculating

- **P(A and B) = P(both event A occurs and event B occurs)**

beginning with the multiplication rule for independent events.

Using a Venn diagram, we can visualize "A and B," which is represented by the overlap between events A and B:



Probability Rule Six (The Multiplication Rule for Independent Events):

- **If A and B are two INDEPENDENT events, then P(A and B) = P(A) * P(B).**

Comment:

- When dealing with probability **rules**, the word **"and"** will always be associated with the operation of **multiplication**; hence the name of this rule, "The Multiplication Rule."

EXAMPLE:

Recall the blood type example:

Blood type	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

Two people are selected simultaneously and at random from all people in the United States.

What is the probability that both have blood type O?

- Let O1= "person 1 has blood type O" and
- O2= "person 2 has blood type O"

We need to **find P(O1 and O2)**

Since they were chosen simultaneously and at random, the blood type of one has no effect on the blood type of the other. Therefore, O1 and O2 are independent, and we may apply Rule 6:

- **P(O1 and O2) = P(O1) * P(O2) = 0.44 * 0.44 = 0.1936.**

Did I Get This?: Probability Rule Six

Comments:

- We now have an Addition Rule that says

P(A or B) = P(A) + P(B) for **disjoint** events,

and a Multiplication Rule that says

P(A and B) = P(A) * P(B) for **independent** events.

The purpose of this comment is to point out the magnitude of P(A or B) and of P(A and B) relative to either one of the individual probabilities.

Since probabilities are never negative, the probability of one event **or** another is always **at least as large as either of the individual probabilities**.

Since probabilities are never more than 1, the probability of one event **and** another generally involves multiplying numbers that are less than 1, therefore **can never be more than either of the individual probabilities**.

Here is an example:

EXAMPLE:

Consider the event A that a randomly chosen person has blood type A.

Modify it to a more general event — that a randomly chosen person has blood type A or B — and the probability increases.

Modify it to a more specific (or restrictive) event — that not just one randomly chosen person has blood type A, but that out of two simultaneously randomly chosen people, person 1 will have type A and person 2 will have type B — and the probability decreases.

It is important to mention this in order to root out a common misconception.

- The word “and” is associated in our minds with “adding more stuff.” Therefore, some students **incorrectly**think that P(A and B) should be larger than either one of the individual probabilities, while it is actually smaller, since it is a more specific (restrictive) event.
- Also, the word “or” is associated in our minds with “having to choose between” or “losing something,” and therefore some students **incorrectly** think that P(A or B) should be smaller than either one of the individual probabilities, while it is actually larger, since it is a more general event.

Practically, you can use this comment to check yourself when solving problems.

For example, if you solve a problem that involves “or,” and the resulting probability is smaller than either one of the individual probabilities, then you know you have made a mistake somewhere.

Did I Get This?: Comparing P(A and B) to P(A or B)

Comment:

- Probability rule six can be used as a test to see if two events are independent or not.
- If you can easily find P(A), P(B), and P(A and B) using logic or are provided these values, then we can test for independent events using the multiplication rule for independent events:

IF P(A)*P(B) = P(A and B) THEN A and B are independent events, otherwise, they are dependent events.

As you've seen, the last three rules that we've introduced (the Complement Rule, the Addition Rules, and the Multiplication Rule for Independent Events) are frequently used in solving problems.

Before we move on to our next rule, here are two comments that will help you use these rules in broader types of problems and more effectively.

Comment:

- As we mentioned before, the Addition Rule for Disjoint events (rule four) can be extended to more than two disjoint events.
- Likewise, the Multiplication Rule for independent events (rule six) can be extended to more than two independent events.
- So if A, B and C are three independent events, for example, then P(A and B and C) = P(A) * P(B) * P(C).
- These extensions are quite straightforward, as long as you remember that "or" requires us to add, while "and" requires us to multiply.

✔ EXAMPLE:

Three people are chosen simultaneously and at random.

What is the probability that all three have blood type B?

We'll use the usual notation of B1, B2 and B3 for the events that persons 1, 2 and 3 have blood type B, respectively.

We need to find P(B1 and B2 and B3). Let's solve this one together:

> **Learn by Doing:** Extending Probability Rule Six

Here is another example that might be quite surprising.

✔ EXAMPLE:

A fair coin is tossed 10 times. Which of the following two outcomes is more likely?

(a) HHHHHHHHHH

(b) HTTHHTHTTH

> **Learn by Doing:** A Surprising Result using Probability Rule Six?

In fact, they are equally likely. The 10 tosses are independent, so we'll use the Multiplication Rule for Independent Events:

- **P(HHHHHHHHHH) = P(H) * P(H) * … *P(H) = 1/2 * 1/2 *… * 1/2 = $(1/2)^{10}$**
- **P(HTTHHTHTTH) = P(H) * P(T) * … * P(H) = 1/2 * 1/2 *… * 1/2 = $(1/2)^{10}$**

Here is the idea:

Our random experiment here is tossing a coin 10 times.

- You can imagine how huge the sample space is.
- There are actually 1,024 possible outcomes to this experiment, all of which are equally likely.

Therefore,

- while it is true that it is more likely to get an outcome that has 5 heads and 5 tails than an outcome that has only heads

since there is only **one possible outcome** which gives **all heads**

and **many possible outcomes** which give **5 heads and 5 tails**

- if we are comparing 2 **specific outcomes**as we do here, they are **equally likely**.

IMPORTANT Comments:

- **Only** use the multiplication rule for **independent events**, rule six, which says P(A and B) = P(A)P(B) if you are certain the two events are independent.
  - **Probability rule six is ONLY true for independent events.**
- **When finding P(A or B) using the general addition rule: P(A) + P(B) – P(A and B)**,
  - **do NOT use the multiplication rule for independent events to calculate P(A and B), use only logic and counting.**

Conditional Probability (Rule Seven)

Learning Objectives

LO 6.9: Apply logic or probability rules to calculate conditional probabilities, P(A|B), and interpret them in context.

Now we will introduce the concept of **conditional probability**.

The idea here is that the probabilities of certain events may be affected by whether or not other events have occurred.

The term "**conditional**" refers to the fact that we will have **additional conditions, restrictions, or other information** when we are asked to calculate this type of probability.

Let's illustrate this idea with a simple example:

EXAMPLE:

All the students in a certain high school were surveyed, then classified according to gender and whether they had either of their ears pierced:

	Pierced	Not Pierced	Total
Male	36	144	180
Female	288	32	320
Total	324	176	500

(Note that this is a two-way table of counts that was first introduced when we talked about the relationship between two categorical variables.

It is not surprising that we are using it again in this example, since we indeed have two categorical variables here:

- **Gender:**M or F (in our notation, "not M")
- **Pierced:**Yes or No

Suppose a student is selected at random from the school.

- Let **M**and **not M** denote the events of being male and female, respectively,
- and **E**and **not E** denote the events of having ears pierced or not, respectively.

What is the probability that the student has either of their ears pierced?

Since a student is chosen at random from the group of 500 students, out of which 324 are pierced,

- **P(E) = 324/500 = 0.648**

What is the probability that the student is male?

Since a student is chosen at random from the group of 500 students, out of which 180 are male,

- **P(M) = 180/500 = 0.36.**

What is the probability that the student is male and has ear(s) pierced?

Since a student is chosen at random from the group of 500 students out of which 36 are male and have their ear(s) pierced,

- **P(M and E) = 36/500 = 0.072**

Now something new:

Given that the student that was chosen is male, what is the probability that he has one or both ears pierced?

At this point, new notation is required, to express the probability of a certain event given that another event holds.

We will write

- "**the probability of having either ear pierced (E), given that a student is male (M)**"
- as **P(E | M).**

A word about this new notation:

- The event whose probability we seek (in this case E) is written first,
- the **vertical line** stands for the word "**given**" or "**conditioned on**,"
- and the event that is given (in this case M) is written after the "|" sign.

We call this probability the

- **conditional probabilityof having either ear pierced, given that a student is male**:
- it assesses the probability of having pierced ears under the condition of being male.

Now to find the probability, we observe that **choosing from only the males** in the school essentially **alters the sample space** from all students in the school **to all male students in the school**.

The **total number of possible outcomes** is **no longer 500**, but **has changed to 180**.

Out of those 180 males, 36 have ear(s) pierced, and thus:

- **P(E | M) = 36/180 = 0.20.**

A good visual illustration of this conditional probability is provided by the two-way table:

	Pierced	Not Pierced	Total
Male	36	144	180
Female	288	32	320
Total	324	176	500

which shows us that conditional probability in this example is the same as the conditional percents we calculated back in section 1. In the above visual illustration, it is clear we are calculating a row percent.

✔ EXAMPLE:

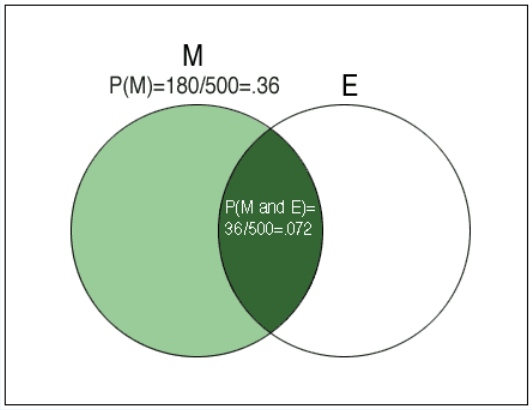
Consider the piercing example, where the following two-way table is given,

	Pierced	Not Pierced	Total
Male	36	144	180
Female	288	32	320
Total	324	176	500

Recall also that M represents the event of being a male ("not M" represents being a female), and E represents the event of having one or both ears pierced.

Did I Get This?: Conditional Probability

Another way to visualize conditional probability is using a Venn diagram:



In both the two-way table and the Venn diagram,

- the reduced sample space (comprised of only males) is shaded light green,
- and within this sample space, the event of interest (having ears pierced) is shaded darker green.

The two-way table illustrates the idea via counts, while the Venn diagram converts the counts to probabilities, which are presented as regions rather than cells.

We may work with counts, as presented in the two-way table, to write

- **P(E | M) = 36/180.**

Or we can work with probabilities, as presented in the Venn diagram, by writing

- **P(E | M) = (36/500) / (180/500).**

We will want, however, to write our formal expression for conditional probabilities in terms of other, ordinary, probabilities and therefore the definition of conditional probability will grow out of the Venn diagram.

Notice that

- **P(E | M) = (36/500) / (180/500) = P(M and E) / P(M).**

Probability Rule Seven (Conditional Probability Rule):

- **The conditional probability of event B, given event A, is P(B | A) = P(A and B) / P(A)**

Comments:

- Note that when we evaluate the conditional probability, we always divide by the probability of the given event. The probability of both goes in the numerator.
- The above formula holds as long as P(A) > 0, since we cannot divide by 0. In other words, we should not seek the probability of an event given that an impossible event has occurred.

Let's see how we can use this formula in practice:

✔ EXAMPLE:

On the "Information for the Patient" label of a certain antidepressant, it is claimed that based on some clinical trials,

- there is a 14% chance of experiencing sleeping problems known as insomnia (denote this event by **I**),
- there is a 26% chance of experiencing headache (denote this event by **H**),
- and there is a 5% chance of experiencing both side effects (**I and H**).

(a) Suppose that the patient experiences insomnia; what is the probability that the patient will also experience headache?

Since we know (or it is **given**) that the patient experienced **insomnia**, we are looking for **P(H | I)**.

According to the definition of conditional probability:

- **P(H | I) = P(H and I) / P(I) = 0.05/0.14 = 0.357.**

(b) Suppose the drug induces headache in a patient; what is the probability that it also induces insomnia?

Here, we are given that the patient experienced headache, so we are looking for P(I | H).

Using the definition

- **P(I | H) = P(I and H) / P(H) = 0.05/0.26 = 0.1923.**

Comment:

- Note that the answers to (a) and (b) above are different.
- **In general, P(A | B) does not equal P(B | A).** We'll come back and illustrate this point later.

Now that we have introduced conditional probability, try the interactive demonstration below which uses a Venn diagram to illustrate the basic probabilities we have been discussing.

Now you can investigate the conditional probabilities as well.

Interactive Applet: Conditional Probability

Independent Events (Part 2)

Learning Objectives

LO 6.7: Determine whether two events are independent or dependent and justify your conclusion.

As we saw in the Exploratory Data Analysis section, whenever a situation involves more than one variable, it is generally of interest to determine whether or not the variables are related.

In probability, we talk about **independent events**, and earlier we said that two events A and B are **independent** if event A occurring **does not affect** the probability that event B will occur.

Now that we've introduced conditional probability, we can formalize the definition of independence of events and develop four simple ways to check whether two events are independent or not.

We will introduce these "**independence checks**" using examples, and then summarize.

EXAMPLE:

Consider again the two-way table for all 500 students in a particular high school, classified according to gender and whether or not they have one or both ears pierced.

	Pierced	Not Pierced	Total
Male	36	144	180
Female	288	32	320
Total	324	176	500

Would you expect those two variables to be related?

- That is, would you expect having pierced ears to depend on whether the student is male or female?
- Or, to put it yet another way, would knowing a student's gender affect the probability that the student's ears are pierced?

To answer this, we may compare the overall probability of having pierced ears to the conditional probability of having pierced ears, given that a student is male.

Our intuition would tell us that the latter should be lower:

- male students tend not to have their ears pierced, whereas female students do.

Indeed, for students in general, the probability of having pierced ears (event E) is

- **P(E) = 324/500 = 0.648.**

But the probability of having pierced ears given that a student is male is only

- **P(E | M) = 36/180 = 0.20.**

As we anticipated, P(E | M) is lower than P(E).

The probability of a student having pierced ears changes (in this case, gets lower) when we know that the student is male, and therefore the events E and M are **dependent**.

Remember, if E and M were independent, knowing or not knowing that the student is male would not have made a difference … but it did.

The previous example illustrates that **one method for determining whether two events are independent is to compare P(B | A) and P(B)**.

- If the two are **equal**(i.e., knowing or not knowing whether A has occurred has no effect on the probability of B occurring) then the two events are **independent**.
- Otherwise, if the **probability changes**depending on whether we know that A has occurred or not, then the two events are **not independent**.

Similarly, using the same reasoning, we can **compare P(A | B) and P(A).**

EXAMPLE:

Recall the side effects activity (from the bottom of the page Basic Probability Rules.).

On the "Information for the Patient" label of a certain antidepressant, it is claimed that based on some clinical trials,

- there is a 14% chance of experiencing sleeping problems known as insomnia (denote this event by **I**),
- there is a 26% chance of experiencing headache (denote this event by **H**),
- and there is a 5% chance of experiencing both side effects (**I and H**).

Are the two side effects independent of each other?

To check whether the two side effects are independent, let's **compare P(H | I) and P(H).**

In the previous part of this section, we found that

- **P(H | I)**= P(H and I) / P(I) = 0.05/0.14 = **0.357,**
- while **P(H) = 0.26.**

Knowing that a patient experienced insomnia increases the likelihood that he/she will also experience headache from 0.26 to 0.357.

The conclusion therefore is that the two side effects are not independent, they are **dependent**.

Alternatively, we could have **compared P(I | H) to P(I).**

- **P(I) = 0.14**,
- and previously we found that **P(I | H)**= P(I and H) / P(H) = 0.05/0.26 = **0.1923,**

Again, since the two are **not equal**, we can conclude that the two side effects I and H are **dependent**.

Comment:

- Recall the pierced ears example. We checked the independence of the events M (being a male) and E (having pierced ears) by comparing P(E) to P(E | M).

An alternative method of checking for dependence would be to compare P(E | M) with P(E | not M) [same as P(E | F)].

In our case, P(E | M) = 36/180 = 0.2, while P(E | not M) = 288/320 = 0.9, and since the two are very different, we can say that the events E and M are not independent.

In general, another method for checking the independence of events A and B is to compare **P(B | A) and P(B | not A)**.

In other words, two events are independent if the probability of one event does not change whether we know that the other event has occurred or we know that the other event has not occurred.

It can be shown that **P(B | A) and P(B | not A)** would differ whenever P(B) and P(B | A) differ, so this is another perfectly legitimate way to establish dependence or independence.

Before we establish a general rule for independence, let's consider an example that will illustrate another method that we can use to check whether two events are independent:

EXAMPLE:

A group of 100 college students were surveyed about their gender and whether they had decided on a major.

	Decided	Undecided	Total
Female	27	33	60
Male	18	22	40
Total	45	55	100

Offhand, we wouldn't necessarily have any compelling reason to expect that deciding on a major would depend on a student's gender.

We can check for independence by comparing the overall probability of being decided to the probability of being decided given that a student is female:

- **P(D) = 45/100 = 0.45 and P(D | F) = 27/60 = 0.45.**

The fact that the two are equal tells us that, as we might expect, deciding on a major is independent of gender.

Now let's approach the issue of independence in a different way: first, we may note that the overall probability of being decided is 45/100 = 0.45.

	Decided	Undecided	Total
Female	27	33	60
Male	18	22	40
Total	45	55	100

And the overall probability of being female is 60/100 = 0.60.

	Decided	Undecided	Total
Female	27	33	60
Male	18	22	40
Total	45	55	100

If being decided is independent of gender, then 45% of the 60% of the class who are female should have a decided major;

in other words, the probability of being female and decided should equal the probability of being female multiplied by the probability of being decided.

If the events F and D are independent, we should have P(F and D) = P(F) * P(D).

In fact, **P(F and D) = 27/100 = 0.27 = P(F) * P(D) = 0.45 * 0.60**.

This confirms our alternate verification of independence.

In general, another method for checking the independence of events A and B is to

- **compare P(A and B) to P(A) * P(B).**
- If the two are **equal**, then A and B are **independent**, **otherwise** the two are **not independent**.

Let's summarize all the possible methods we've seen for checking the independence of events in one rule:

Tests for Independent Events: Two events A and B are independent if any one of the following hold:

- **P(B | A) = P(B)**
- **P(A | B) = P(A)**
- **P(B | A) = P(B | not A)**
- **P(A and B) = P(A) * P(B)**

Comment:

- These various equalities turn out to be equivalent, so that if one equality holds, all are equal, and if one equality does not hold, all are not equal. (This is the case for the same reason that knowing one of the values P(A and B), P(A and not B), P(not A and B), or P(not A and not B), along with P(A) and P(B), allows you to determine the remaining cells of a two-way probability table.)
- Therefore, in order to check whether events A and B are independent or not, it is sufficient to check only whether one of the four equalities holds — whichever is easiest for you.

The purpose of the next activity is to practice checking the independence of two events using the four different possible methods that we've provided, and see that all of them will lead us to the same conclusion, regardless of which of the four methods we use.

Learn by Doing: Tests for Independent Events

General Multiplication Rule (Rule Eight)

Learning Objectives

LO 6.10: Use the general multiplication rule to calculate P(A and B) for any events A and B.

Now that we have an understanding of conditional probabilities and can express them with concise notation, and have a more formal understanding of what it means for two events to be independent, we can finally establish the **General Multiplication Rule**, a formal rule for finding **P(A and B)** that applies to any two events, whether they are independent or dependent.

We begin with an example that contrasts P(A and B) for independent and dependent cases.

EXAMPLE:

Suppose you pick two cards at random from four cards consisting of one of each suit: **club, diamond, heart, and spade**, where the first card is replaced before the second card is picked.

What is the probability of picking a club and then a diamond?

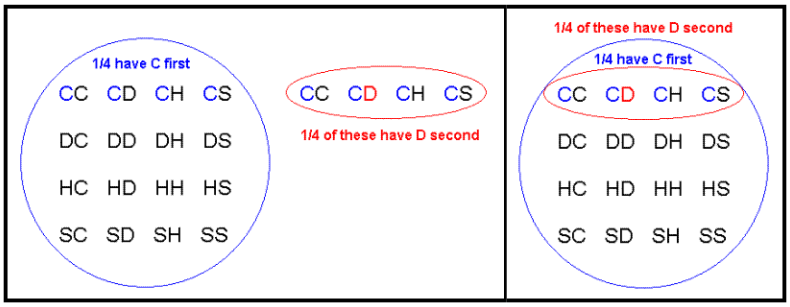
Because the sampling is done with replacement, whether or not a diamond is picked on the second selection is independent of whether or not a club has been picked on the first selection.

Rule 6, the multiplication rule for independent events, tells us that:

- **P(C1 and D2) = P(C1) * P(D2) = 1/4 * 1/4 = 1/16.**

Here we denote the event "club picked on first selection" as C1 and the event "diamond picked on second selection" as D2.

The display below shows that 1/4 of the time we'll pick a club first, and of these times, 1/4 will result in a diamond on the second pick: 1/4 * 1/4 = 1/16 of the selections will have a club first and then a diamond.



EXAMPLE:

Suppose you pick two cards at random from four cards consisting of one of each suit: **club, diamond, heart, and spade**, without replacing the first card before the second card is picked.

What is the probability of picking a club and then a diamond?

The probability in this case is **not** 1/4 * 1/4 = 1/16.

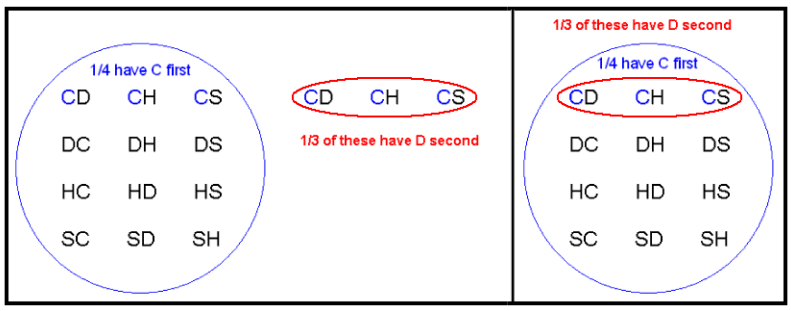
- Because the sampling is done without replacement, so whether or not a diamond is picked on the second selection **does** depend on what was picked on the first selection.
- For instance, if a diamond was picked on the first selection, the probability of another diamond is zero!
- As in the example above, 1/4 of the time we'll pick a club first.
- But since the club has been removed, 1/3 of these selections with a club first will have a diamond second.

The probability of a club and then a diamond is **1/4*1/3=1/12.**

- This is the probability of getting a club first, multiplied by the probability of getting a diamond second, given that a club was picked first.

Using the notation of conditional probabilities, we can write

- **P(C1 and D2) = P(C1) * P(D2 | C1) = 1/4 * 1/3 = 1/12.**



For independent events A and B, we had the rule P(A and B) = P(A) * P(B).

Due to independence, to find the probability of A and B, we could multiply the probability of A by the simple probability of B, because the occurrence of A would have no effect on the probability of B occurring.

Now, for events A and B that may be dependent, to find the probability of A and B, we multiply the probability of A by the **conditional probability of B**, taking into account that A has occurred.

Thus, our general multiplication rule is stated as follows:

General Multiplication Rule – Probability Rule Eight:

- **For any two events A and B, P(A and B) = P(A) * P(B | A)**

Comments:

1. Note that although the motivation for this rule was to find P(A and B) when A and B are not independent, this rule is general in the sense that if A and B happen to be **independent**, then P(B | A) = P(B) is true, and we're back to Rule 6 — the Multiplication Rule for Independent Events: P(A and B) = P(A) * P(B).

---

2. The General Multiplication Rule is just the definition of conditional probability in disguise. Recall the definition of conditional probability: **P(B | A) = P(A and B) / P(A)** Let's isolate P(A and B) by multiplying both sides of the equation by P(A), and we get: **P(A and B) = P(A) * P(B | A)**. That's it … this is the General Multiplication Rule.

---

3. The General Multiplication Rule is useful when two events, A and B, occur in stages, first A and then B (like the selection of the two cards in the previous example). Thinking about it this way makes the General Multiplication Rule very intuitive. For both A and B to occur you first need A to occur (which happens with probability P(A)), **and** then you need B to occur, knowing that A has already occurred (which happens with probability P(B | A)).

Did I Get This?: The General Multiplication Rule

Let's look at another, more realistic example:

EXAMPLE:

In a certain region, one in every thousand people (0.001) is infected by the HIV virus that causes AIDS.

- Tests for presence of the virus are fairly accurate but not perfect.
- If someone actually has HIV, the probability of testing positive is 0.95.

Let **H** denote the event of having HIV, and **T** the event of testing positive.

(a) Express the information that is given in the problem in terms of the events H and T.

- "one in every thousand people (0.001) of all individuals are infected with HIV" → **P(H) = 0.001**
- "If someone actually has HIV, the probability of testing positive is 0.95" → **P(T | H) =0.95**

(b) Use the General Multiplication Rule to find the probability that someone chosen at random from the population has HIV and tests positive.

- **P(H and T)= P(H) * P(T | H) = 0.001*0.95 = 0.00095.**

(c) If someone has HIV, what is the probability of testing negative? Here we need to find P(not T | H).

- The Complement Rule works with conditional probabilities as long as we **condition on the same event**, therefore:
- **P(not T | H)= 1 – P(T | H) = 1 – 0.95 = 0.05.**

The purpose of the next activity is to give you guided practice in expressing information in terms of conditional probabilities, and in using the General Multiplication Rule.

Learn by Doing: Conditional Probability and the General Multiplication Rule

Let's Summarize

This section introduced you to the fundamental concepts of **independent events** and **conditional probability** — the probability of an event given that another event has occurred.

We saw that sometimes the knowledge that another event has occurred has no impact on the probability (when the two events are **independent**), and sometimes it does (when the two events are not independent).

We further discussed the idea of independence and discussed different ways to check whether two events are independent or not.

Understanding the concept of conditional probability also allowed us to introduce our final probability rule, the **General Multiplication Rule**.

The General Multiplication Rule tells us how to find P(A and B) when A and B are not necessarily independent.

Introduction to Probability

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Video: Probability Introduction (7:41)

Now that we understand how probability fits into the Big Picture as a key element behind statistical inference, we are ready to learn more about it. Our first goal is to introduce some fundamental terminology (the language) and notation that is used when discussing probability.

Probability is Not Always Intuitive

Although most of the probability calculations we will conduct will be rather intuitive due to their simplicity, we start with two fun examples that will illustrate the interesting and sometimes complex nature of probability.

Often, relying only on our intuition is not enough to determine probability, so we'll need some tools to work with, which is exactly what we'll study in this section.

Caution

For the next two examples, do not be concerned with the solution of the problem. Only how the answers to probability questions are not always easy to believe or determine.

Here is the first of two motivating examples:

EXAMPLE: The "Let's Make a Deal" Paradox

"Let's Make a Deal" was the name of a popular television game show, which first aired in the 1960s. The "Let's Make a Deal" Paradox is named after that show. In the show, the contestant had to choose between three doors. One of the doors had a big prize behind it such as a car or a lot of cash, and the other two were empty. (Actually, for entertainment's sake, each of the other two doors had some stupid gift behind it, like a goat or a chicken, but we'll refer to them here as empty.)

The contestant had to choose one of the three doors, but instead of revealing the chosen door, the host revealed one of the two unchosen doors to be empty. At this point of the game, there were two unopened doors (one of which had the prize behind it) — the door that the contestant had originally chosen and the remaining unchosen door.

The contestant was given the option either to **stay** with the door that he or she had initially chosen, **or switch** to the other door.

What do you think the contestant should do, **stay or switch**? What do you think is the probability that you will win the big prize if you stay? What about if you switch?

In order for you to gain a feel for this game, you can play it a few times using an applet.

Interactive Applet: Let's Make a Deal

Now, what do you think a contestant should do?

Learn By Doing: Let's Make a Deal

The intuition of most people is that the chance of winning is equal whether we stay or switch — that there is a 50-50 chance of winning with either selection. This, however, is not the case.

Actually, there is a 67% chance — or a probability of 2/3 (2 out of three) — of winning by switching, and only a 33% chance — or a probability of 1/3 (1 out of 3) — of winning by staying with the door that was originally chosen.

This means that a contestant is twice as likely to win if he/she switches to the unchosen door. Isn't this a bit counterintuitive and confusing? Most people think so, when they are first faced with this problem.

We will now try to explain this paradox to you in two different ways:

> **Video:** Let's Make a Deal (Explanation #1) (1:10)

If you are still not convinced (or even if you are), here is a different way of explaining the paradox:

> **Video:** Let's Make a Deal (Explanation #2) (1:37)

If this example still did not persuade you that probability is not always intuitive, the next example should definitely do the trick.

EXAMPLE: The Birthday Problem

Suppose that you are at a party with 59 other people (for a total of 60). What are the chances (or, what is the probability) that at least 2 of the 60 guests share the same birthday?

To clarify, by "share the same birthday," we mean that 2 people were born on the same date, not necessarily in the same year. Also, for the sake of simplicity, ignore leap years, and assume that there are 365 days in each year.

> **Learn By Doing:** Birthday Problem

Indeed, there is a 99.4% chance that at least 2 of the 60 guests share the same birthday. In other words, it is **almost certain** that at least 2 of the guests share the same birthday. This is very counterintuitive.

Unlike the "Let's Make a Deal" example, for this scenario, we don't really have a good step-by-step explanation that will give you insight into this surprising answer.

From these two examples, (maybe) you have seen that your original hunches cannot always be counted upon to give you correct predictions of probabilities.

We won't think any more about these examples as they are from the "harder" end of the complexity spectrum but hopefully they have motivated you to learn more about probability and you do not need to be convinced of their solution to continue!

In general, **probability is not always intuitive.**

Need a Laugh?

Watch this (funny) video which has an excellent point about "how probability DOES NOT work": clip from the Daily Show with Jon Stewart about the Large Hadron Collider (5:58).

It is possible viewers in other countries may not be able to view the clip from this source. You may or may not be able to find it online through searching. Here is the transcript summary I sometimes use in class to get the point across (it isn't quite as funny but I think you can still figure out what is wrong here):

- John Oliver: So, roughly speaking, what are the chances that the world is going to be destroyed? (by the large hadron collider) One-in-a-million? One-in-a-billion?
- **Walter: Well, the best we can say right now is about a one-in-two chance.**
- John Oliver: 50-50?
- **Walter: Yeah, 50-50… It's a chance; it's a 50-50 chance.**
- John Oliver: You keep coming back to this 50-50 thing, it's weird Walter.
- **Walter: Well, if you have something that can happen and something that won't necessarily happen, it's going to either happen or it's going to not happen. And, so, it's … the best guess is 1 in 2.**

- John Oliver: I'm not sure that's how probability works, Walter.

And … John Oliver is correct! :-)

What is Probability?

Learning Objectives

LO 6.4: Relate the probability of an event to the likelihood of this event occurring.

Eventually we will need to develop a more formal approach to probability, but we will begin with an informal discussion of what probability is.

Probability is a mathematical description of randomness and uncertainty. It is a way to measure or quantify uncertainty. Another way to think about probability is that it is the official name for "chance."

Probability is the Likelihood of Something Happening

One way to think of probability is that it is the **likelihood** that something will occur.

Probability is used to answer the following types of questions:

- What is the chance that it will rain tomorrow?
- What is the chance that a stock will go up in price?
- What is the chance that I will have a heart attack?
- What is the chance that I will live longer than 70 years?
- What is the likelihood that when rolling a pair of dice, I will roll doubles?
- What is the probability that I will win the lottery?
- What is the probability that I will become diabetic?

Each of these examples has some uncertainty. For some, the chances are quite good, so the probability would be quite high. For others, the chances are not very good, so the probability is quite low (especially winning the lottery).

Certainly, the chance of rain is different each day, and is higher during some seasons. Your chance of having a heart attack, or of living longer than 70 years, depends on things like your current age, your family history, and your lifestyle. However, you could use your intuition to predict some of those probabilities fairly accurately, while others you might have no instinct about at all.

Notation

We think you will agree that the word **probability** is a bit long to include in equations, graphs and charts, so it is customary to use some simplified notation instead of the entire word.

If we wish to indicate "the probability it will rain tomorrow," we use the notation "P(rain tomorrow)." We can abbreviate the probability of anything. If we let **A** represent what we wish to find the probability of, then **P(A)** would represent that probability.

We can think of "**A**" as an "***event***."

NOTATION	MEANING
P(win lottery)	the probability that a person who has a lottery ticket will win that lottery
P(A)	the probability that event A will occur
P(B)	the probability that event B will occur

PRINCIPLE: The "probability" of an event tells us how likely it is that the event will occur.

What values can the probability of an event take, and what does the value tell us about the likelihood of the event occurring?

Video

Video: Basic Properties of Probability (0:53)

Did I Get This?: Basic Properties of Probability

PRINCIPLE: The probability that an event will occur is between 0 and 1 or $0 \le P(A) \le 1$.

Many people prefer to express probability in percentages. Since all probabilities are decimals, each can be changed to an equivalent percentage. Thus, the latest principle is equivalent to saying, **"The chance that an event will occur is between 0% and 100%**."

Probabilities can be determined in two fundamental ways. Keep reading to find out what they are.

Determining Probability

There are 2 fundamental ways in which we can determine probability:

- Theoretical (also known as Classical)
- Empirical (also known as Observational)

Classical methods are used for games of chance, such as flipping coins, rolling dice, spinning spinners, roulette wheels, or lotteries.

The probabilities in this case are determined by the game (or scenario) itself and are often found relatively easily using logic and/or probability rules.

Although we will not focus on this type of probability in this course, we will mention a few examples to get you thinking about probability and how it works.

EXAMPLE: Flipping a Coin

A coin has two sides; we usually call them "heads" and "tails."

For a "fair" coin (one that is not unevenly weighted, and does not have identical images on both sides) the chances that a "flip" will result in either side facing up are equally likely.

Thus, P(heads) = P(tails) = 1/2 or 0.5.

Letting **H** represent "heads," we can abbreviate the probability: **P(H) = 0.5.**

Classical probabilities can also be used for more realistic and useful situations.

A practical use of a coin flip would be for you and your roommate to decide randomly who will go pick up the pizza you ordered for dinner. A common expression is "Let's flip for it." This is because a coin can be used to make a random choice with two options. Many sporting events begin with a coin flip to determine which side of the field or court each team will play on, or which team will have control of the ball first.

EXAMPLE: Rolling a Fair Die

Each traditional (cube-shaped) die has six sides, marked in dots with the numbers 1 through 6.

On a "fair" die, these numbers are equally likely to end up face-up when the die is rolled.

Thus, P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = **1/6 or about 0.167.**

Here, again, is a practical use of classical probability.

Suppose six people go out to dinner. You want to randomly decide who will pick up the check and pay for everyone. Again, the P(each person) = 1/6.

EXAMPLE: Spinners

This particular spinner has three colors, but each color is not equally likely to be the result of a spin, since the portions are not the same size.

Since the blue is half of the spinner, P(blue) = 1/2. The red and yellow make up the other half of the spinner and are the same size. Thus, P(red) = P(yellow) = 1/4.

Suppose there are 2 freshmen, 1 sophomore, and one junior in a study group. You want to select one person. The P(F) = 2/4 = 1/2; P(S) = 1/4; and P(J) = 1/4, just like the spinner.

EXAMPLE: Selecting Students

Suppose we had three students and wished to select one of them randomly. To do this you might have each person write his/her name on a (same-sized) piece of paper, then put the three papers in a hat, and select one paper from the hat without looking.

Since we are selecting randomly, each is equally likely to be chosen. Thus, each has a probability of 1/3 of being chosen.

A slightly more complicated, but more interesting, probability question would be to propose selecting 2 of the students pictured above, and ask, “What is the probability that the two students selected will be different genders?”

We will now shift our discussion to empirical ways to determine probabilities.

A Question

A single flip of a coin has an uncertain outcome. So, every time a coin is flipped, the outcome of that flip is unknown until the flip occurs.

However, if you flip a fair coin over and over again, would you expect P(H) to be exactly 0.5? In other words, would you expect there to be the same number of results of “heads” as there are “tails”?

The following activity will allow you to discover the answer.

Learn By Doing: Empirical Probability #1

The above Learn by Doing activity was our first example of the second way of determining probability: Empirical (Observational) methods. In the activity, we determined that the probability of getting the result “heads” is 0.5 by flipping a fair coin many, many times.

A Second Question

After doing this experiment, an important question naturally comes to mind. **How would we know if the coin was not fair?** Certainly, classical probability methods would never be able to answer this question. In addition, classical methods could never tell us the actual P(H). The only way to answer this question is to perform another experiment.

The next activity will allow you to do just that.

> **Learn By Doing:** Empirical Probability #2

So, these types of experiments can verify classical probabilities and they can also determine when games of chance are not following **fair** practices. However, their real importance is to answer probability questions that arise when we are faced with a situation that does not follow any pattern and cannot be predetermined. In reality, most of the probabilities of interest to us fit the latter description.

To Summarize So Far

1. Probability is a way of quantifying uncertainty.
2. We are interested in the probability of an event — the likelihood of the event occurring.
3. The probability of an event ranges from 0 to 1. The closer the probability is to 0, the less likely the event is to occur. The closer the probability is to 1, the more likely the event is to occur.
4. There are two ways to determine probability: Theoretical (Classical) and Empirical (Observational).
5. Theoretical methods use the nature of the situation to determine probabilities.
6. Empirical methods use a series of trials that produce outcomes that cannot be predicted in advance (hence the uncertainty).

Relative Frequency

Learning Objectives

LO 6.5: Apply the relative frequency approach to estimate the probability of an event.

If we toss a coin, roll a die, or spin a spinner many times, we hardly ever achieve the exact **theoretical** probabilities that we know we should get, but we can get pretty close. When we run a simulation or when we use a random sample and record the results, we are using **empirical** probability. This is often called the **Relative Frequency** definition of probability.

Here is a realistic example where the relative frequency method was used to find the probabilities:

EXAMPLE: Blood Type

Researchers discovered at the beginning of the 20th century that human blood comes in various types (A, B, AB, and O), and that some types are more common than others. How could researchers determine the probability of a particular blood type, say O?

Just looking at one or two or a handful of people would not be very helpful in determining the overall chance that a randomly chosen person would have blood type O. But sampling many people at random, and finding the relative frequency of blood type O occurring, provides an adequate estimate.

For example, it is now well known that the probability of blood type O among white people in the United States is 0.45. This was found by sampling many (say, 100,000) white people in the country, finding that roughly 45,000 of them had blood type O, and then using the relative frequency: 45,000 / 100,000 = 0.45 as the estimate for the probability for the event **"having blood type O."**

(Comment: Note that there are racial and ethnic differences in the probabilities of blood types. For example, the probability of blood type O among black people in the United States is 0.49, and the probability that a randomly chosen Japanese person has blood type O is only 0.3).

Let's review the relative frequency method for finding probabilities:

To estimate the probability of event A, written P(A), we may repeat the random experiment many times and count the number of times event A occurs. Then P(A) is estimated by the ratio of the number of times A occurs to the number of repetitions, which is called the **relative frequency of event A**.

$$\text{Relative Frequency of Event A} = \frac{\text{number of times A occurred}}{\text{total number of repetitions}}$$

Did I Get This?: Relative Frequency

Learn By Doing: Relative Frequency

So, we've seen how the relative frequency idea works, and hopefully the activities have convinced you that the relative frequency of an event does indeed approach the theoretical probability of that event as the number of repetitions increases. This is called the **Law of Large Numbers**.

The Law of Large Numbers states that as the number of trials increases, the relative frequency becomes the actual probability. So, using this law, as the number of trials increases, the empirical probability gets closer and closer to the theoretical probability.

PRINCIPLE: Law of Large Numbers – The actual (or true) probability of an event (A) is estimated by the relative frequency with which the event occurs in a long series of trials.

Interactive Applet: Law of Large Numbers

Comments:

1. Note that the relative frequency approach provides only an estimate of the probability of an event. However, we can control how good this estimate is by the number of times we repeat the random experiment. The more repetitions that are performed, the closer the **relative frequency** gets to the **true probability** of the event.

---

2. One interesting question would be: "How many times do I need to repeat the random experiment in order for the relative frequency to be, say, within 0.001 of the actual probability of the event?" We will come back to that question in the **inference** section.

---

3. A pedagogical comment: We've introduced relative frequency here in a more practical approach, as a method for estimating the probability of an event. More traditionally, relative frequency is not presented as a method, but as a definition:

Relative Frequency: (Definition) The probability of an event (A) is the relative frequency with which the event occurs in a long series of trials.

4. There are many situations of interest in which physical circumstances do not make the probability obvious. In fact, most of the time it is impossible to find the theoretical probability, and we must use empirical probabilities instead.

Let's Summarize

Probability is a way of quantifying uncertainty. In this section, we defined **probability** as the **likelihood** or chance that something will occur and introduced the basic **notation** of probability such as P(win lottery).

You have seen that all probabilities are values between 0 and 1, where an event with no chance of occurring has a probability of 0 and an event which will always occur has a probability of 1.

We have discussed the two primary methods of calculating probabilities

- Theoretical or **Classical** Probability: uses the nature of the situation to determine probabilities

- **Empirical** or Observational Probability: uses a series of trials that produce outcomes that cannot be predicted in advance (hence the uncertainty)

In our course we will focus on Empirical probability and will often calculate probabilities from a sample using **relative frequencies**.

This is useful in practice since the **Law of Large Numbers** allows us to estimate the actual (or true) probability of an event by the relative frequency with which the event occurs in a long series of trials. We can collect this information as data and we can analyze this data using statistics.

Summary (Unit 3)

Video

Video: Live Examples – Calcium Oxalate Crystals (40:18 Total)

Video

Video: Live Examples – Diabetes (10:33 Total)

This summary provides a quick recap of the material you've learned in the probability unit so far. Please note that this summary does not provide complete coverage of the material, but just lists the main points. We therefore recommend that you use this summary only as a checklist or a review before going on to the next unit, or before an exam.

General Remarks

- Probability is a discipline by itself. In the context of the big picture of this course, probability is used to quantify the imperfection associated with drawing conclusions about the entire population based only on a random sample drawn from it.
- The probability of an event can be as low as 0 (when the event is impossible) and as high as 1 (when the event is certain).
- In some cases, the only way to find the probability of an event of interest is by repeating the random experiment many times and using the relative frequency approach.
- When all the possible outcomes of a random experiment are equally likely, the probability of an event is the fraction of outcomes which satisfy it.
- There are many applications of probability in the health sciences including sensitivity, specificity, predictive value positive, predictive value negative, relative risk, odds ratios, to name a few.

Probability Principles

Probability principles help us find the probability of events of certain types:

- **The Complement Rule**, P(not A) = 1 – P(A), is especially useful for finding events of the type "at least one of …"
- To find the probability of **events of the type "A or B"** (interpreted as A occurs or B occurs or both), we use the General Addition Rule: P(A or B) = P(A) + P(B) – P(A and B).In the special case when A and B are disjoint (cannot happen together; P(A and B) = 0) the Addition Rule reduces to: P(A or B) = P(A) + P(B).
- To find the probability of **events of the type "A and B"** (interpreted as both A and B occur), we use the General Multiplication Rule: P(A and B) = P(A) * P(B | A). In the special case when A and B are independent (the occurrence of one event has no effect on the probability of the other occurring; P(B | A) = P(B)) the Multiplication Rule **reduces** to: P(A and B) = P(A) * P(B).
- Both **restricted versions** of the addition rule (for disjoint events) and the multiplication rule (for independent events) **can be extended** to more than two events.
- P(B | A), the **conditional probability** of event B occurring given that event A has occurred, can be viewed as a reduction of the sample space S to event A. The conditional probability, then, is the fraction of event A where B occurs as well, P(B | A) = P(A and B) / P(A).
- Be sure to follow reasonable rounding rules for probability, including enough significant digits and avoiding any rounding in intermediate steps.

(Optional) Outside Reading: Little Handbook – Probability (≈ 1000 words)

CHAPTER OVERVIEW

Unit 3B: Random Variables

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Video: Unit 3B Random Variables (10:00)

Introduction

In the remaining sections in Unit 3 we will begin to make the **connection** between **probability** and **statistics** so that we can apply these concepts in the final Unit on statistical inference.

These concepts **bridge the gap** between the **mathematics of descriptive statistics** and **probability** and **true "Inferential Statistics"** where we will formalize **statistical hypothesis tests**.

In other words, the topics in Unit 3B provide the **mathematical background** and **concepts** that will be needed for our **study of inferential statistics**.

In the previous sections we learned principles and tools that help us find probabilities of events in general.

Now that we've become proficient at doing that, we'll talk about **random variables**.

Just like any other variable, random variables can take on multiple values.

> What differentiates **random variables** from other variables is that the **values** for these variables are **determined by a random trial, random sample, or simulation.**

The probabilities for the values can be determined by theoretical or observational means.

Such probabilities play a vital role in the theory behind statistical inference, our ultimate goal in this course.

Random Variables

Learning Objectives

LO 6.11: Distinguish between discrete and continuous random variables

We first discussed variables in the Exploratory Data Analysis portion of the course. A variable is a characteristic of an individual.

We also made an important distinction between **categorical variables**, whose values are groups or categories (and an individual can be placed into one of them), and **quantitative variables**, which have numerical values for which arithmetic operations make sense.

In the previous sections, we focused mostly on events which arise when there is a categorical variable in the background: blood type, pierced ears (yes/no), gender, on time delivery (yes/no), side effect (yes/no), etc.

Now we will begin to consider quantitative variables that arise when a random experiment is performed. We will need to define this new type of variable.

> A **random variable** assigns a unique numerical value to the outcome of a random experiment.

A random variable can be thought of as a function that associates exactly one of the possible numerical outcomes to each trial of a random experiment. However, that number can be the same for many of the trials.

Before we go any further, here are some simple examples:

EXAMPLE: Theortical

Consider the random experiment of flipping a coin twice.

- The sample space of possible outcomes is S = { HH, HT, TH, TT }.

Now, let's **define the variable X to be the number of tails** that the random experiment will produce.

- If the outcome is HH, we have no tails, so the value for **X is 0**.
- If the outcome is HT, we got one tail, so the value for **X is 1**.
- If the outcome is TH, we again got one tail, so the value for **X is 1**.
- Lastly, if the outcome is TT, we got two tails, so the value for **X is 2**.

As the definition suggests, **X is a quantitative variable that takes the possible values of 0, 1, or 2.**

It is **random** because we do not know which of the three values the variable will eventually take.

We can ask questions like:

- What is the probability that X will be 2? In other words, what is the probability of getting 2 tails?
- What is the probability that X will be at least 1? In other words, what is the probability of getting at least 1 tail?

As you can see, random variables are not really a new thing, but just a different way to look at the same problem.

Note that if we had tossed a coin three times, the possible values for the number of tails would be 0, 1, 2, or 3. In general, if we toss a coin "n" times, the possible number of tails would be 0, 1, 2, 3, … , or n.

EXAMPLE: Observational

Consider getting data from a random sample on the number of ears in which a person wears one or more earrings.

We **define the variable X to be the number of ears** in which a randomly selected person wears an earring.

- If the selected person does not wear any earrings, then **X = 0**.
- If the selected person wears earrings in either the left or the right ear, then **X = 1**.
- If the selected person wears earrings in both ears, then **X = 2**.

As the definition suggests, **X is a quantitative variable which takes the possible values of 0, 1, or 2**.

We can ask questions like:

- What is the probability that a randomly selected person will have earrings in both ears?
- What is the probability that a randomly selected person will not be wearing any earrings in either ear?

NOTE… We identified the first example as **theoretical** and the second as **observational**.

Let's discuss the distinction.

- To answer probability questions about a theoretical situation, we only need the principles of probability.
- However, if we have an observational situation, the only way to answer probability questions is to use the relative frequency we obtain from a random sample.

Here is a different type of example:

EXAMPLE: Lightweight Boxer

Assume we choose a lightweight male boxer at random and record his exact weight.

According to the boxing rules, a lightweight male boxer must weigh between 130 and 135 pounds, so the sample space here is

- S = { All the numbers in the interval 130-135 }.

Note that we can't list all the possible outcomes here!

We'll **define X to be the weight of the boxer** again, as the definition suggests, **X is a quantitative variable whose value is the result of our random experiment**.

Here **X can take any value between 130 and 135**.

We can ask questions like:

- What is the probability that X will be more than 132? In other words, what is the probability that the boxer will weigh more than 132 pounds?
- What is the probability that X will be between 131 and 133? In other words, what is the probability that the boxer weighs between 131 and 133 pounds?

What is the difference between the random variables in these examples? Let's see:

- They all arise from a random experiment (tossing a coin twice, choosing a person at random, choosing a lightweight boxer at random).
- They are all quantitative (number of tails, number of ears, weight).

Where they differ is in the type of possible values they can take:

- In the first two examples, X has three distinct possible values: 0, 1, and 2. You can list them.
- In contrast, in the third example, X takes any value in the interval 130-135, and thus the possible values of X cover an infinite range of possibilities, and cannot be listed.

Types of Random Variables

A random variable like the one in the first two examples, whose possible values are a list of distinct values, is called a **discrete random variable**.

A random variable like the one in the third example, that can take any value in an interval, is called a **continuous random variable**.

The main distinction between these two types of random variables is that,

- although they can both take on a potentially infinite number of values,
- for **discrete** random variables there is always a **GAP** between any two possible values
- whereas for **continuous** random variables there are no gaps in the range of possible values – it can take on any value in an interval; our precision in measurement is only limited by our level of technology in taking that measurement.

Just as the distinction between categorical and quantitative variables was important in Exploratory Data Analysis, the distinction between discrete and continuous random variables is important here, as each one gets a different treatment when it comes to calculating probabilities and other quantities of interest.

Before we go any further, a few observations about the nature of discrete and continuous random variables should be mentioned.

Comments:

- Sometimes, continuous random variables are "rounded" and are therefore "in a discrete disguise." For example:
  - time spent watching TV in a week, rounded to the nearest hour (or minute)
  - outside temperature, to the nearest degree
  - a person's weight, to the nearest pound.

Even though they "look like" discrete variables, these are still continuous random variables, and we will in most cases treat them as such.

- On the other hand, there are some variables which are discrete in nature, but take so many distinct possible values that it will be much easier to treat them as continuous rather than discrete.
  - the IQ of a randomly chosen person
  - the SAT score of a randomly chosen student
  - the annual salary of a randomly chosen CEO, whether rounded to the nearest dollar or the nearest cent

- Sometimes we have a discrete random variable but do not know the extent of its possible values.
  - For example: How many accidents will occur in a particular intersection this month?
  - We may know from previously collected data that this number is from 0-5. But, 6, 7, or more accidents could be possible.
- A good rule of thumb is that **discrete** random variables are things we **count**, while **continuous** random variables are things we **measure**.
  - We counted the number of tails and the number of ears with earrings. These were discrete random variables.
  - We measured the weight of the lightweight boxer. This was a continuous random variable.

Often we can have a subject matter for which we can collect data that could involve a discrete or a continuous random variable, depending on the information we wish to know.

✔ EXAMPLE: Soft Drinks

Suppose we want to know **how many days per week you drink a soft drink**.

- The sample space would be S = { 0, 1, 2, 3, 4, 5, 6, 7 }.
- There are a finite number of values for this variable.
- This would be a **discrete** random variable.

Instead, suppose we want to know **how many ounces of soft drinks you consume per week**.

- Even if we round to the nearest ounce, the answer is a measurement.
- Thus, this would be a **continuous** random variable.

✔ EXAMPLE: x-bar

Suppose we are interested in the weights of all males.

- We take a random sample and get the mean for that sample, namely x-bar.
- We then take another random sample (with the same sample size) and get another x-bar.
- We would expect the values of the x-bars from these two samples to be different, but pretty close in value.
- Each time we take a sample we'll get a different x-bar.
- We will take lots of samples and thus get many x-bar values.

The value of x-bar from these repeated samples is a **random variable**.

Since it can take on any value within an interval of possible male weights it is a **continuous** random variable.

Did I Get This?: Random Variables

We devote a great deal of attention to random variables, since random variables and the probabilities that are associated with them play a vital role in the theory behind statistical inference, our ultimate goal in this course.

We'll start with discrete random variables, including a discussion of binomial random variables and then move on to continuous random variables where we will formalize our understanding of the normal distribution.

Binomial Random Variables

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Review:

- Basic Probability Rules
- Conditional Probability and Independence

Video: Binomial Random Variables (12:52)

So far, in our discussion about discrete random variables, we have been introduced to:

1. The probability distribution, which tells us which values a variable takes, and how often it takes them.
2. The mean of the random variable, which tells us the long-run average value that the random variable takes.
3. The standard deviation of the random variable, which tells us a typical (or long-run average) distance between the mean of the random variable and the values it takes.

We will now introduce a special class of discrete random variables that are very common, because as you'll see, they will come up in many situations – **binomial random variables.**

Here's how we'll present this material.

- First, we'll explain what kind of random experiments give rise to a binomial random variable, and how the binomial random variable is defined in those types of experiments.
- We'll then present the probability distribution of the binomial random variable, which will be presented as a formula, and explain why the formula makes sense.
- We'll conclude our discussion by presenting the mean and standard deviation of the binomial random variable.

As we just mentioned, we'll start by describing what kind of random experiments give rise to a binomial random variable. We'll call this type of random experiment a "binomial experiment."

Binomial Experiment

Learning Objectives

LO 6.14: When appropriate, apply the binomial model to find probabilities.

Binomial experiments are random experiments that consist of a fixed number of repeated trials, like tossing a coin 10 times, randomly choosing 10 people, rolling a die 5 times, etc.

These trials, however, need to be **independent** in the sense that the outcome in one trial has no effect on the outcome in other trials.

In each of these repeated trials there is one outcome that is of interest to us (we call this outcome "success"), and each of the trials is identical in the sense that the probability that the trial will end in a "success" is the same in each of the trials.

So for example, if our experiment is tossing a coin 10 times, and we are interested in the outcome "heads" (our "success"), then this will be a binomial experiment, since the 10 trials are independent, and the probability of success is 1/2 in each of the 10 trials.

Let's summarize and give more examples.

The **requirements** for a random experiment to be a **binomial experiment** are:

- a fixed number (n) of trials
- each trial must be independent of the others
- each trial has just two possible outcomes, called "**success**" (the outcome of interest) and "**failure**"

- there is a constant **probability (p) of success** for each trial, the complement of which is the **probability (1 – p) of failure, sometimes denoted as q = (1 – p)**

In binomial random experiments, the number of successes in n trials is random.

It can be as low as 0, if all the trials end up in failure, or as high as n, if all n trials end in success.

The random variable X that represents the number of successes in those n trials is called a **binomial** random variable, and is determined by the values of n and p. We say, "X is binomial with n = … and p = …"

EXAMPLE: Random Experiments (Binomial or Not?)

Let's consider a few random experiments.

In each of them, we'll decide whether the random variable is binomial. If it is, we'll determine the values for n and p. If it isn't, we'll explain why not.

Example A:

A fair coin is flipped 20 times; X represents the number of heads.

X is binomial with n = 20 and p = 0.5.

Example B:

You roll a fair die 50 times; X is the number of times you get a six.

X is binomial with n = 50 and p = 1/6.

Example C:

Roll a fair die repeatedly; X is the number of rolls it takes to get a six.

X is not binomial, because the number of trials is not fixed.

Example D:

Draw 3 cards at random, one after the other, **without replacement**, from a set of 4 cards consisting of one club, one diamond, one heart, and one spade; X is the number of diamonds selected.

X is not binomial, because the selections are not independent. (The probability (p) of success is not constant, because it is affected by previous selections.)

Example E:

Draw 3 cards at random, one after the other, **with replacement**, from a set of 4 cards consisting of one club, one diamond, one heart, and one spade; X is the number of diamonds selected. Sampling with replacement ensures independence.

X is binomial with n = 3 and p = 1/4

Example F:

Approximately 1 in every 20 children has a certain disease. Let X be the number of children with the disease out of a random sample of 100 children. Although the children are sampled without replacement, it is assumed that we are sampling from such a vast population that the selections are virtually independent.

X is binomial with n = 100 and p = 1/20 = 0.05.

Example G:

The probability of having blood type B is 0.1. Choose 4 people at random; X is the number with blood type B.

X is binomial with n = 4 and p = 0.1.

Example H:

A student answers 10 quiz questions completely at random; the first five are true/false, the second five are multiple choice, with four options each. X represents the number of correct answers.

X is not binomial, because p changes from 1/2 to 1/4.

Comments:

- **Example D** above was not binomial because sampling without replacement resulted in dependent selections.
  - In particular, the probability of the second card being a diamond is very dependent on whether or not the first card was a diamond:
  - the probability is 0 if the first card was a diamond, 1/3 if the first card was not a diamond.
- In contrast, **Example E** was binomial because sampling with replacement resulted in independent selections:
  - the probability of any of the 3 cards being a diamond is 1/4 no matter what the previous selections have been.
- On the other hand, when you take a relatively small random sample of subjects from a large population, even though the sampling is without replacement, we can assume independence because the mathematical effect of removing one individual from a very large population on the next selection is negligible.
  - For example, in **Example F**, we sampled 100 children out of the population of all children.
  - Even though we sampled the children without replacement, whether one child has the disease or not really has no effect on whether another child has the disease or not.
  - The same is true for **Example (G.)**.

Did I Get This?: Binomial or Not?

Binomial Probability Distribution – Using Probability Rules

Now that we understand what a binomial random variable is, and when it arises, it's time to discuss its probability distribution. We'll start with a simple example and then generalize to a formula.

EXAMPLE: Deck of Cards

Consider a regular deck of 52 cards, in which there are 13 cards of each suit: hearts, diamonds, clubs and spades. We select 3 cards at random **with replacement**. Let X be the number of diamond cards we got (out of the 3).

We have 3 trials here, and they are independent (since the selection is with replacement). The outcome of each trial can be either success (diamond) or failure (not diamond), and the probability of success is 1/4 in each of the trials.

X, then, is binomial with n = 3 and p = 1/4.

Let's build the probability distribution of X as we did in the chapter on probability distributions. Recall that we begin with a table in which we:

- record all possible outcomes in 3 selections, where each selection may result in success (a diamond, D) or failure (a non-diamond, N).
- find the value of X that corresponds to each outcome.
- use simple probability principles to find the probability of each outcome.

Outcome	X	Probability
NNN	0	3/4*3/4*3/4
NND	1	3/4*3/4*1/4
NDN	1	3/4*1/4*3/4
DNN	1	1/4*3/4*3/4
NDD	2	3/4*1/4*1/4
DND	2	1/4*3/4*1/4
DDN	2	1/4*1/4*3/4
DDD	3	1/4*1/4*1/4

With the help of the addition principle, we condense the information in this table to construct the actual probability distribution table:

X	0	1	2	3
P(X=x)	$1(3/4)^3$	$3(1/4)^1(3/4)^2$	$3(1/4)^2(3/4)^1$	$1(1/4)^3$

In order to establish a general formula for the probability that a binomial random variable X takes any given value x, we will look for patterns in the above distribution. From the way we constructed this probability distribution, we know that, in general:

$P(X = x) =$ Number of possible outcomes with x successes out of 3 $*$ Probability of each of the outcomes that has x successes out of 3

Let's start with the second part, the probability that there will be x successes out of 3, where the probability of success is 1/4.

Notice that the fractions multiplied in each case are for the probability of x successes (where each success has a probability of p = 1/4) and the remaining (3 – x) failures (where each failure has probability of 1 – p = 3/4).

X	0	1	2	3
P(X=x)	$1\,(1/4)^0(3/4)^{3-0}$	$3(1/4)^1(3/4)^{3-1}$	$3(1/4)^2(3/4)^{3-2}$	$1(1/4)^3(3/4)^{3-3}$

So in general:

Probability of each of the outcomes that has x successes out of 3 $= (1/4)^x * (3/4)^{3-x}$

Let's move on to talk about the number of possible outcomes with x successes out of three. Here it is harder to see the pattern, so we'll give the following mathematical result.

Counting Outcomes

Consider a random experiment that consists of n trials, each one ending up in either success or failure. The number of possible outcomes in the sample space that have exactly k successes out of n is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The notation on the left is often read as "n choose k." Note that n! is read "n factorial" and is defined to be the product 1 * 2 * 3 * … * n. 0! is defined to be 1.

EXAMPLE: Ear Piercings

You choose 12 male college students at random and record whether they have any ear piercings (success) or not. There are many possible outcomes to this experiment (actually, 4,096 of them!).

In how many of the possible outcomes of this experiment are there exactly 8 successes (students who have at least one ear pierced)?

There is no way that we would start listing all these possible outcomes. The result above comes to our rescue.

The result says that in an experiment like this, where you repeat a trial n times (in our case, we repeat it n = 12 times, once for each student we choose), the number of possible outcomes with exactly 8 successes (out of 12) is:

$$\frac{12!}{8!(12-8)!} = \frac{1*2*3*\cdots*12}{(1*2*3*\cdots*8)(1*2*3*4)} = 495$$

Did I Get This?: Counting Outcomes

EXAMPLE: Card Revisited

Let's go back to our example, in which we have n = 3 trials (selecting 3 cards). We saw that there were 3 possible outcomes with exactly 2 successes out of 3. The result confirms this since:

$$\frac{3!}{2!(3-2)!} = \frac{1*2*3}{(1*2)(1)} = \frac{6}{2} = 3$$

In general, then

Number of possible outcomes with x successes out of 3 $= \frac{3!}{x!\,(3-x)!}$

Putting it all together, we get that the probability distribution of X, which is binomial with n = 3 and p = 1/4 i

$$P(X=x) = \frac{3!}{x!(3-x)!}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{3-x} \quad x = 0, 1, 2, 3$$

In general, the number of ways to get x successes (and n – x failures) in n trials is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Therefore, the probability of x successes (and n – x failures) in n trials, where the probability of success in each trial is p (and the probability of failure is 1 – p) is equal to the number of outcomes in which there are x successes out of n trials, times the probability of x successes, times the probability of n – x failures:

Binomial Probability Formula for P(X = x)

$$P(X=x) = \frac{n!}{x!(n-x)!}p^{x}(1-p)^{(n-x)}$$

where x may take any value 0, 1, … , n.

Let's look at another example:

EXAMPLE: Blood Type A

The probability of having blood type A is 0.4. Choose 4 people at random and let X be the number with blood type A.

X is a binomial random variable with n = 4 and p = 0.4.

As a review, let's first find the probability distribution of X the long way: construct an interim table of all possible outcomes in S, the corresponding values of X, and probabilities. Then construct the probability distribution table for X.

S	X	Probability
NNNN	0	$.4^{0}.6^{4}$
NNNA	1	$.4^{1}.6^{3}$
NNAN	1	$.4^{1}.6^{3}$
NANN	1	$.4^{1}.6^{3}$
ANNN	1	$.4^{1}.6^{3}$
NNAA	2	$.4^{2}.6^{2}$
NANA	2	$.4^{2}.6^{2}$
NAAN	2	$.4^{2}.6^{2}$
ANNA	2	$.4^{2}.6^{2}$
ANAN	2	$.4^{2}.6^{2}$
AANN	2	$.4^{2}.6^{2}$
NAAA	3	$.4^{3}.6^{1}$
ANAA	3	$.4^{3}.6^{1}$
AANA	3	$.4^{3}.6^{1}$
AAAN	3	$.4^{3}.6^{1}$
AAAA	4	$.4^{4}.6^{0}$

As usual, the addition rule lets us combine probabilities for each possible value of X:

X	Probability
0	(1) $.4^0.6^4$ =.1296
1	(4) $.4^1.6^3$ =.3456
2	(6) $.4^2.6^2$ =.3456
3	(4) $.4^3.6^1$ =.1536
4	(1) $.4^4.6^0$ =.0256

Now let's apply the formula for the probability distribution of a binomial random variable, and see that by using it, we get exactly what we got the long way.

Recall that the general formula for the probability distribution of a binomial random variable with n trials and probability of success p is:

$$P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)} \text{ for x} = 0,1,2,3,\ldots,\text{n}$$

In our case, X is a binomial random variable with n = 4 and p = 0.4, so its probability distribution is:

$$P(X=x)=\frac{4!}{x!(4-x)!}(0.4)^x(0.6)^{4-x} \text{ for x} = 0,1,2,3,4$$

Let's use this formula to find P(X = 2) and see that we get exactly what we got before.

$$P(X=2)=\frac{4!}{2!(4-2)!}(0.4)^2(0.6)^{4-2}=\frac{1^*2^*3^*4}{(1^*2)\,(1^*2)}(0.4)^2(0.6)^2=0.3456$$

Learn by Doing: Binomial Probabilities (Using Online Calculator)

Now let's look at some truly practical applications of binomial random variables.

EXAMPLE: Airline Flights

Past studies have shown that 90% of the booked passengers actually arrive for a flight. Suppose that a small shuttle plane has 45 seats. We will assume that passengers arrive independently of each other. (This assumption is not really accurate, since not all people travel alone, but we'll use it for the purposes of our experiment).

Many times airlines "*overbook*" flights. This means that the airline sells more tickets than there are seats on the plane. This is due to the fact that sometimes passengers don't show up, and the plane must be flown with empty seats. However, if they do overbook, they run the risk of having more passengers than seats. So, some passengers may be unhappy. They also have the extra expense of putting those passengers on another flight and possibly supplying lodging.

With these risks in mind, the airline decides to sell more than 45 tickets. If they wish to keep the probability of having more than 45 passengers show up to get on the flight to less than 0.05, how many tickets should they sell?

This is a binomial random variable that represents the number of passengers that show up for the flight. It has p = 0.90, and n to be determined.

Suppose the airline sells 50 tickets. Now we have n = 50 and p = 0.90. We want to know P(X > 45), which is 1 – P(X ≤ 45) = 1 – 0.57 or 0.43. Obviously, all the details of this calculation were not shown, since a statistical technology package was used to calculate the answer. This is certainly more than 0.05, so the airline must sell fewer seats.

If we reduce the number of tickets sold, we should be able to reduce this probability. We have calculated the probabilities in the following table:

# tickets sold	P(X > 45)
50	45)" class="lt-stats-31283">45)" class=" ">0.43
49	45)" class="lt-stats-31283">45)" class=" ">0.26
48	45)" class="lt-stats-31283">45)" class=" ">0.13

47	45)" class="lt-stats-31283">45)" class=" ">0.04
46	45)" class="lt-stats-31283">45)" class=" ">0.008

From this table, we can see that by selling 47 tickets, the airline can reduce the probability that it will have more passengers show up than there are seats to less than 5%.

Note: For practice in finding binomial probabilities, you may wish to verify one or more of the results from the table above.

Learn by Doing: Binomial Application

Mean and Standard Deviation of the Binomial Random Variable

Learning Objectives

LO 6.15: Find the mean, variance, and standard deviation of a binomial random variable.

Now that we understand how to find probabilities associated with a random variable X which is binomial, using either its probability distribution formula or software, we are ready to talk about the mean and standard deviation of a binomial random variable. Let's start with an example:

EXAMPLE: Blood Type B - Mean

Overall, the proportion of people with blood type B is 0.1. In other words, roughly 10% of the population has blood type B.

Suppose we sample 120 people at random. On average, how many would you expect to have blood type B?

The answer, 12, seems obvious; automatically, you'd multiply the number of people, 120, by the probability of blood type B, 0.1.

This suggests the general formula for finding the mean of a binomial random variable:

Claim:

If X is binomial with parameters n and p, then the **mean** or **expected value** of X is:

$$\mu_X = np$$

Although the formula for mean is quite intuitive, it is not at all obvious what the variance and standard deviation should be. It turns out that:

Claim:

The ideal gas law is easy to remember and apply in solving problems, as long as you get the **proper values a**If X is binomial with parameters n and p, then the **variance** and **standard deviation** of X are:

$$\sigma_X^2 = np(1-p)$$
$$\sigma_X = \sqrt{np(1-p)}$$

Comments:

- The binomial mean and variance are special cases of our general formulas for the mean and variance of any random variable. Clearly it is much simpler to use the "shortcut" formulas presented above than it would be to calculate the mean and variance or standard deviation from scratch.
- Remember, these "shortcut" formulas only hold in cases where you have a binomial random variable.

EXAMPLE: Blood Type B - Standard Deviation

Suppose we sample 120 people at random. The number with blood type B should be about 12, give or take how many? In other words, what is the standard deviation of the number X who have blood type B?

Since n = 120 and p = 0.1,

$$\sigma_X^2 = 120(0.1)(1-0.1) = 10.8; \quad \sigma_X = \sqrt{10.8} \approx 3.3$$

In a random sample of 120 people, we should expect there to be about 12 with blood type B, give or take about 3.3.

Did I Get This?: Binomial Distribution

Before we move on to continuous random variables, let's investigate the shape of binomial distributions.

Learn by Doing: Shapes of Binomial Distributions

Continuous Random Variables

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Video: Continuous Random Variables (3:59)

In the previous section, we discussed discrete random variables: random variables whose possible values are a list of distinct numbers. We talked about their probability distributions, means, and standard deviations.

We are now moving on to discuss continuous random variables: random variables which can take any value in an interval, so that all of their possible values cannot be listed (such as height, weight, temperature, time, etc.)

As it turns out, most of the methods for dealing with continuous random variables require a higher mathematical level than we needed to deal with discrete random variables. For the most part, the calculation of probabilities associated with a continuous random variable, and its mean and standard deviation, requires knowledge of calculus, and is beyond the scope of this course.

What we will do in this part is discuss the idea behind the probability distribution of a continuous random variable, and show how calculations involving such variables become quite complicated very fast!

We'll then move on to a special class of continuous random variables – normal random variables. Normal random variables are very common, and play a very important role in statistical inference.

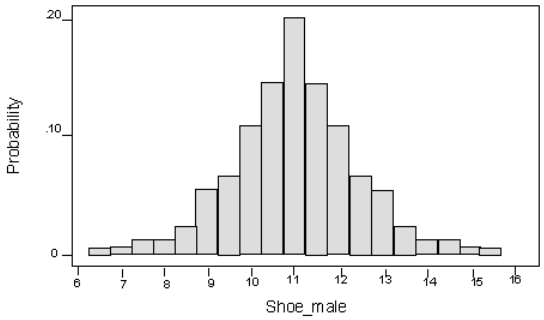
We'll finish this section by presenting an important connection between the binomial random variable (the special discrete random variable that we presented earlier) and the normal random variable (the special continuous random variable that we'll present here).

The Probability Distribution of a Continuous Random Variable

Learning Objectives

LO 6.16: Explain how a density function is used to find probabilities involving continuous random variables.

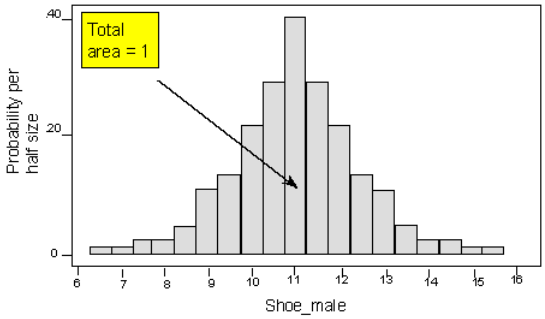
In order to shift our focus from discrete to continuous random variables, let us first consider the probability histogram below for the shoe size of adult males. Let X represent these shoe sizes. Thus, X is a discrete random variable, since shoe sizes can only take whole and half number values, nothing in between.



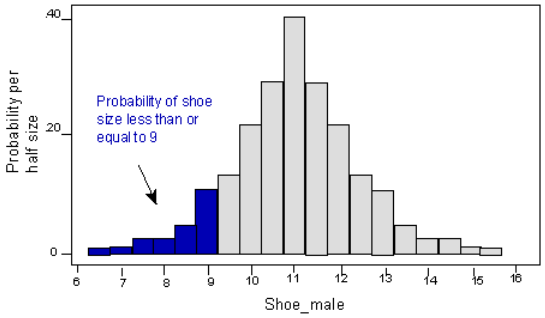
Recall that in all of the previous probability histograms we've seen, the X-values were whole numbers. Thus, the width of each bar was 1. The height of each bar was the same as the probability for its corresponding X-value. Due to the principle that states the sum of probabilities of all possible outcomes in the sample space must be 1, the **heights** of all the rectangles in the histogram must sum to 1. This meant that the area was also 1.

This histogram uses half-sizes. We wish to keep the area = 1, but we still want the horizontal scale to represent half-sizes. Therefore, we must adjust the vertical scale of the histogram. As is, the total area of the histogram rectangles would be .50 times the sum of the probabilities, since the width of each bar is .50. Thus, the area is .50(1) = .50. If we double the vertical scale, the

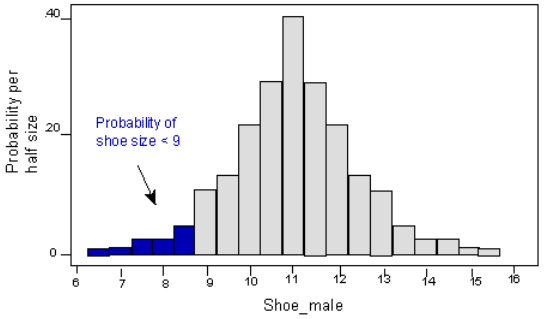
area will double and be 1, just like we want. This means we are changing the vertical scale from “Probability” to “Probability per half size.” The shape and the horizontal scale remain unchanged.



Now we can tell the probability of shoe size taking a value in any interval, just by finding the area of the rectangles over that interval. For instance, the area of the rectangles up to and including 9 shows the probability of having a shoe size less than or equal to 9.



Recall that for a discrete random variable like shoe size, the probability is affected by whether we want strict inequality or not. For example, the area -and corresponding probability – is reduced if we only consider shoe sizes strictly less than 9:



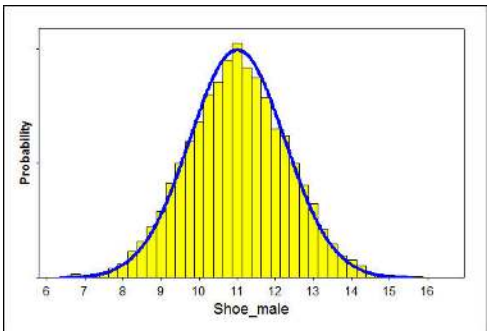
Did I Get This?: Probability for Discrete Random Variables

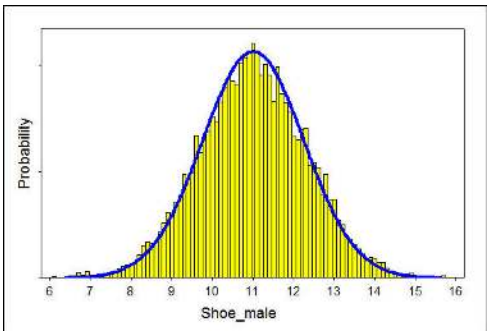
Transition to Continuous Random Variables

Now we are going to be making the transition from **discrete** to **continuous** random variables. Recall that continuous random variables represent measurements and can take on any value within an interval.

For our shoe size example, this would mean measuring shoe sizes in smaller units, such as tenths, or hundredths. As the number of intervals increases, the width of the bars becomes narrower and narrower, and the graph approaches a smooth curve.

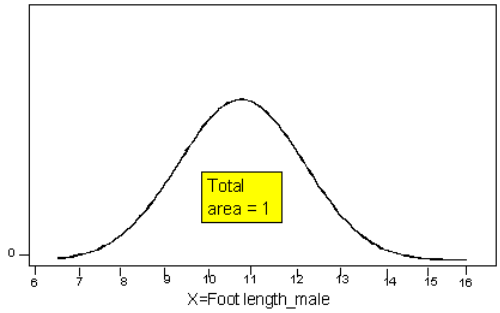
To illustrate this, the following graphs represent two steps in this process of narrowing the widths of the intervals. Specifically, the interval widths are 0.25 and 0.10.





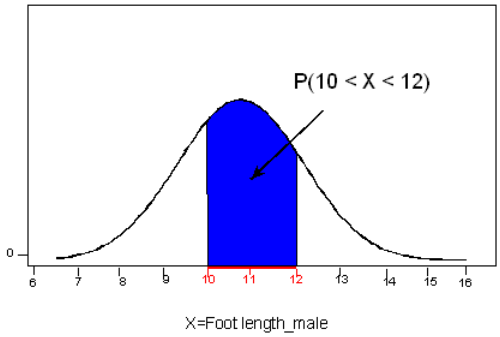
We'll use these smooth curves to represent the probability distributions of continuous random variables. This idea will be discussed in more detail on the next page.

Now consider another random variable X = foot length of adult males. Unlike shoe size, this variable is not limited to distinct, separate values, because foot lengths can take any value over a **continuous** range of possibilities, so we cannot present this variable with a probability histogram or a table. The probability distribution of foot length (or any other continuous random variable) can be represented by a smooth curve called a **probability density curve.**

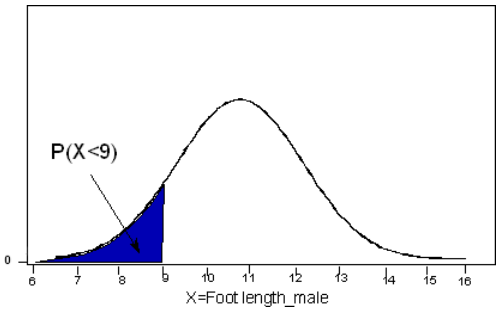


Like the modified probability histogram above, the total area under the density curve equals 1, and the curve represents probabilities by area.

The probability that X gets values in any interval is represented by the area above this interval and below the density curve. In our foot length example, if our interval of interest is between 10 and 12 (marked in red below), and we would like to know P(10 < X < 12), the probability that a randomly chosen male has a foot length anywhere between 10 and 12 inches, we'll have to find the area above our interval of interest (10,12) and below our density curve, shaded in blue:



If, for example, we are interested in P(X < 9), the probability that a randomly chosen male has a foot length of less than 9 inches, we'll have to find the area shaded in blue below:



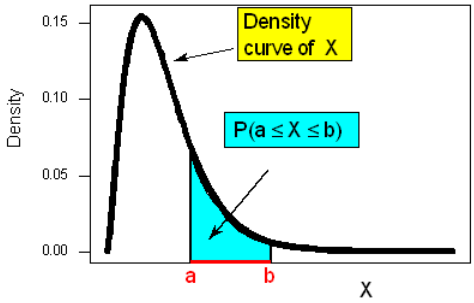
Comments:

- We have seen that for a **discrete** random variable like shoe size, whether we have a strict inequality or not **does matter** when solving for probabilities. In contrast, for a **continuous** random variable like foot length, the probability of a foot length of less than or equal to 9 will be the same as the probability of a foot length of strictly less than 9. In other words, $P(X < 9) = P(X \leq 9)$.Visually, in terms of our density curve, the area under the curve up to and including a certain point is the same as the area up to and excluding the point, because there is no area over a single point. Conceptually, because a continuous random variable has infinitely many possible values, technically the probability of any single value occurring is zero!
- It should be clear now why the total area under any probability density curve must be 1. The total area under the curve represents P(X gets a value in the interval of its possible values). Clearly, according to the rules of probability this must be 1, or always true.
- Density curves, like probability histograms, may have any shape imaginable as long as the total area underneath the curve is 1.

Let's Summarize

The probability distribution of a continuous random variable is represented by a probability density curve.

The probability that X gets a value in any interval of interest is the area above this interval and below the density curve.



Now that we see how probabilities are found for continuous random variables, we understand why it is more complicated than finding probabilities in the discrete case. As anyone who has studied calculus can attest, finding the area under a curve can be difficult. The general approach is to use **integrals.** For those of you who did study calculus, the following should be familiar….

$$P\left(a \leq X \leq b\right) = \left(\text{area between a and b and below the density curve}\right) = \int_a^b f\left(x\right)dx\,,$$

where f(x) represents the density curve.

For those who did not study calculus, don't worry about it. This kind of calculation is definitely beyond the scope of this course.

In this course, we will encounter several important density curves—those for normal random variables, t random variables, chi-square random variables, and F random variables. Normal and t distributions are bell-shaped (single-peaked and symmetric) like the density curve in the foot length example; chi-square and F distributions are single-peaked and skewed right, like in the figure above.

Rather than get bogged down in the calculus of solving for areas under curves, we will find probabilities for the above-mentioned random variables by consulting tables. Also, statistical software automatically provides such probabilities in the appropriate context.

In the next section, we will study in more depth one of those random variables, the normal random variable, and see how we can find probabilities associated with it using software and tables.

Discrete Random Variables

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Review:

- Basic Probability Rules
- Conditional Probability and Independence

Video

Video: Discrete Random Variables (22:40 Total)

We begin with discrete random variables: variables whose possible values are a list of distinct values. In order to decide on some notation, let's look at the coin toss example again:

A fair coin is tossed twice.

- Let the random variable X be the number of tails we get in this random experiment.
- In this case, the possible values that X can assume are
  - 0 (if we get HH),
  - 1 (if get HT or TH),
  - and 2 (if we get TT).

Notation

If we want to find the probability of the event "getting 1 tail," we'll write: **P(X = 1)**

If we want to find the probability of the event "getting 0 tails," we'll write: **P(X = 0)**

In general, we'll write: **P(X = x)** or **P(X = k)** to denote the probability that the **discrete** random variable **X** gets the value **x or k** respectively.

Many students prefer the second notation as keeping track of the difference between X and x can cause confusion.

- Here the X represents the random variable and x or k denote the value of interest in the current problem (0, 1, etc.).
- Note that for the random variables we'll use a capital letter, and for the value we'll use a lowercase letter.

Section Plan

The way this section on discrete random variables is organized is very similar to the way we organized our discussion about one quantitative variable in the Exploratory Data Analysis unit.

It will be separated into four sections.

1. We'll first discuss the probability **distribution** of a discrete random variable, ways to display it, and how to use it in order to find probabilities of interest.
2. We'll then move on to talk about the **mean and standard deviation** of a discrete random variable, which are measures of the center and spread of its distribution.
3. We'll conclude this part by discussing a special and very common class of discrete random variable: the **binomial** random variable.

Probability Distributions

Learning Objectives

LO 6.12: Use the probability distribution for a discrete random variable to find the probability of events of interest.

When we learned how to find probabilities by applying the basic principles, we generally focused on just one particular outcome or event, like the probability of getting exactly one tail when a coin is tossed twice, or the probability of getting a 5 when a die is rolled.

Now that we have mastered the solution of individual probability problems, we'll proceed to look at the big picture by considering all the possible values of a discrete random variable, along with their associated probabilities.

This list of possible values and probabilities is called the **probability distribution** of the random variable.

Comments:

- In the Exploratory Data Analysis unit of this course, we often looked at the distribution of sample values in a quantitative data set. We would display the values with a histogram, and summarize them by reporting their mean.
- In this section, when we look at the probability distribution of a random variable, we consider all its possible values and their overall probabilities of occurrence.
- Thus, we have in mind an entire population of values for a variable. When we display them with a histogram or summarize them with a mean, these are representing a population of values, not a sample.
- The distinction between sample and population is an essential concept in statistics, because an ultimate goal is to draw conclusions about unknown values for a population, based on what is observed in the sample.

In the examples which follow we will sometimes illustrate how the probability distribution is created.

We do this to demonstrate the usefulness of the probability rules we previously discussed and to illustrate clearly how probability distributions can be created.

As we are more focused on data driven methods, you will often be given a probability distribution based upon data as opposed to constructing the theoretical probability distribution based upon flipping coins or similar classical probability experiments.

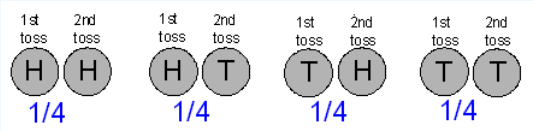
Recall our first example, when we introduced the idea of a random variable. In this example we tossed a coin twice.

EXAMPLE: Flipping a Coin Twice

What is the probability distribution of X, where the random variable X is the number of tails appearing in two tosses of a fair coin?

We first note that since the coin is fair, each of the four outcomes HH, HT, TH, TT in the sample space S is equally likely, and so each has a probability of 1/4.

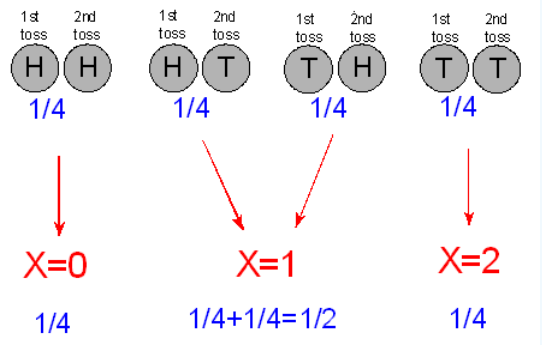
(Alternatively, the multiplication principle can be applied to find the probability of each outcome to be 1/2 * 1/2 = 1/4.)



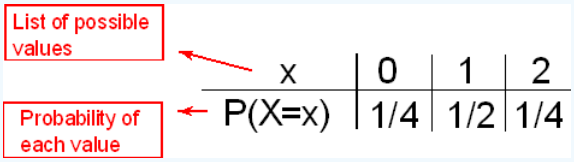
X takes the value 0 only for the outcome **HH**, so the probability that **X = 0 is 1/4.**

X takes the value 1 for outcomes **HT** or **TH**. By the addition principle, the probability that **X = 1 is 1/4 + 1/4 = 1/2.**

Finally, X takes the value 2 only for the outcome **TT**, so the probability that **X = 2 is 1/4**.



The **probability distribution of the random variable X** is easily summarized in a table:



As mentioned before, we write "P(X = x)" to denote "the probability that the random variable X takes the value x."

The way to interpret this table is:

- X takes the values 0, 1, 2 and P(X = 0) = 1/4, P(X = 1) = 1/2, P(X = 2) = 1/4.

Note that events of the type (X = x) are subject to the principles of probability established earlier, and will provide us with a way of systematically exploring the behavior of random variables.

In particular, the first two principles in the context of probability distributions of random variables will now be stated.

The ideal gas law is easy to remember and apply in solving problems, as long as you get the **proper values a**Any **probability distribution** of a **discrete random variable** must satisfy:

$$1.\ 0 \le P(X = x) \le 1$$
$$2.\ \sum_x P(X = x) = 1$$

The probability distribution for two flips of a coin was simple enough to construct at once.

For more complicated random experiments, it is common to first construct a table of all the outcomes and their probabilities, then use the addition principle to condense that information into the actual probability distribution table.

EXAMPLE: Flipping a Coin Three Times

A coin is tossed three times. Let the random variable X be the number of tails.

Find the probability distribution of X.

We'll follow the same reasoning we used in the previous example:

First, we specify the 8 possible outcomes in S, along with the number and the probability of that outcome.

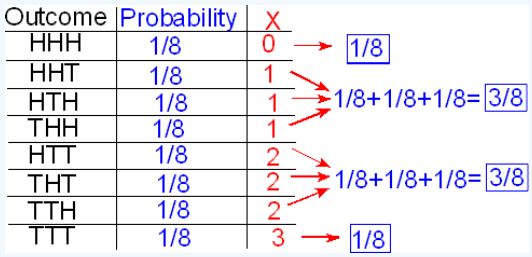
- Because they are all equally likely, each has probability 1/8.
- Alternatively, by the multiplication principle, each particular sequence of three coin faces has probability 1/2 * 1/2 * 1/2 = 1/8.

Then we figure out what the value of X is (number of tails) for each possible outcome.

Outcome	Probability	X
HHH	1/2*1/2*1/2=1/8	0
HHT	1/8	1
HTH	1/8	1
THH	1/8	1
HTT	1/8	2
THT	1/8	2
TTH	1/8	2
TTT	1/8	3

Next, we use the addition principle to assert that

- **P(X = 1) = P(HHT or HTH or THH) = P(HHT) + P(HTH) + P(THH) = 1/8 + 1/8 + 1/8 = 3/8.**
- **Similarly, P(X = 2) = P(HTT or THT or TTH) = 3/8.**



The resulting probability distribution is:

X	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

In the previous two examples, we needed to specify the probability distributions ourselves, based on the physical circumstances of the situation.

In some situations, the probability distribution may be specified with a formula.

Such a formula must be consistent with the constraints imposed by the laws of probability, so that the probability of each outcome must be between 0 and 1, and the probabilities of all possible outcomes together must sum to 1.

We will see this with the binomial distribution.

Probability Histograms

We learned to display the distribution of sample values for a quantitative variable with a histogram in which the horizontal axis represented the range of values in the sample.

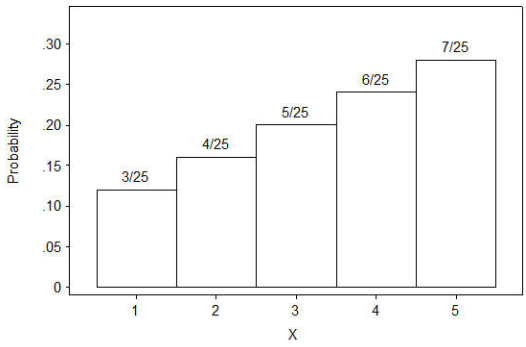
- The vertical axis represented the frequency or relative frequency (sometimes given as a percentage) of sample values occurring in that interval.
- The width of each rectangle in the histogram was an interval, or part of the possible values for the quantitative variable.
- The height of each rectangle was the frequency (or relative frequency) for that interval.

Similarly, we can display the probability distribution of a random variable with a probability histogram.

- The horizontal axis represents the range of all possible values of the random variable
- The vertical axis represents the probabilities of those values.

Here an example of a probability histogram.

(Such probabilities are not always increasing; they just happen to be so in this example).



Area of a Probability Histogram

Notice that each rectangle in the histogram has a width of 1 unit. The height of each rectangle is the probability that it will occur.

Thus, the area of each rectangle is base times height, which for these rectangles is 1 times its probability for each value of X.

> This means that for **probability distributions of discrete random variables**, the sum of the areas of all of the rectangles is the same as the sum of all of the probabilities. **The total area = 1**.

For probability distributions of discrete random variables, this is equivalent to the property that the sum of all of the probabilities must equal 1.

> **Learn by Doing:** Probability Distributions

Finding Probabilities

We've seen how probability distributions are created. Now it's time to use them to find probabilities.

✔ EXAMPLE: Changing Majors

A random sample of graduating seniors was surveyed just before graduation. One question that was asked is:

How many times did you change majors?

The results are displayed in a probability distribution.

x	0	1	2	3	4	5
P(X = x)	.28	.37	.23	.09	.02	.01

Using this probability distribution, we can answer probability questions such as:

What is the probability that a randomly selected senior has changed majors more than once?

This can be written as P(X > 1).

We can find this probability by adding the appropriate individual probabilities in the probability distribution.

- **P(X > 1)**
- **= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)**
- **= 0.23 + 0.09 + 0.02 + 0.01**
- **= 0.35**

As you just saw in this example, we need to pay attention to the wording of the probability question.

The key words that told us which values to use for X are **more than**.

The following will clarify and reinforce the **key words** and their meanings.

Key Words

Let's begin with some everyday situations using **at least** and **at most**.

Suppose someone said to you, "I need you to write **at least 10 pages** for a term paper."

- What does this mean?
- It means that 10 pages is the smallest amount you are going to write.
- In other words, you will write **10 or more**pages for the term paper.
- This would be the same as saying, "**not less than**10 pages."
- So, for example, writing 9 pages would be unacceptable.

On the other hand, suppose you are considering the number of children you will have. You want **at most 3 children**.

- This means that 3 children is the most that you wish to have.
- In other words, you will have **3 or fewer**
- This would be the same as saying, "**not more than**3 children."
- So, for example, you would not want to have 4 children.

The following table gives a list of some key words to know.

Suppose a random variable X had possible values of 0 through 5.

Key Words	Meaning	Symbols	Values for X
more than 2	strictly larger than 2	$X > 2$	3, 4, 5
no more than 2	2 or fewer	$X \leq 2$	0, 1, 2
fewer than 2	strictly smaller than 2	$X < 2$	0, 1
no less than 2	2 or more	$X \geq 2$	2, 3, 4, 5
at least 2	2 or more	$X \geq 2$	2, 3, 4, 5
at most 2	2 or fewer	$X \leq 2$	0, 1, 2
exactly 2	2, no more or no less, only 2	$X = 2$	2

Before we move on to the next section on the means and variances of a probability distribution, let's revisit the changing majors example:

EXAMPLE: Changing Major

x	0	1	2	3	4	5
P(X = x)	.28	.37	.23	.09	.02	.01

Question: Based upon this distribution, do you think it would be unusual to change majors 2 or more times?

Answer:

- **$P(X \geq 2) = 0.35$.**
- So, 35% of the time a student changes majors 2 or more times.
- This means that it is not unusual to do so.

Question: Do you think it would be unusual to change majors 4 or more times?

Answer:

- **$P(X \geq 4) = 0.03$.**
- So, 3% of the time a student changes majors 4 or more times.
- This means that it is fairly unusual to do so.

We can even answer more difficult questions using our probability rules!

Question: What is the probability of changing majors only once given at least one change in major.

Answer:

- **P(X = 1 | X ≥ 1) = P(X = 1 AND X ≥ 1)/P(X ≥ 1)** [using Probability Rule 7]
- **= P(X = 1)/P(X ≥ 1)** [since the only outcome that satisfies both X = 1 and X ≥ 1 is X = 1]
- **= (0.37)/(0.37+0.23+.0.09+0.02+0.01) = 0.37/0.72 = 0.5139.**
- So, among students who change majors, 51% of these students will only change majors one time.

After we learn about means and standard deviations, we will have another way to answer these types of questions.

Mean of a Discrete Random Variable

Learning Objectives

LO 6.13: Find the mean, variance, and standard deviation of a discrete random variable.

In the Exploratory Data Analysis (EDA) section, we displayed the distribution of one quantitative variable with a histogram, and supplemented it with numerical measures of center and spread.

We are doing the same thing here.

- We display the probability distribution of a discrete random variable with a table, formula or histogram.
- And supplement it with numerical measures of the center and spread of the probability distribution.

These measures are the **mean** and **standard deviation** of the **random variable**.

This section will be devoted to introducing these measures. As before, we'll start with the numerical measure of center, the mean. Let's begin by revisiting an example we saw in EDA.

EXAMPLE: World Cup Soccer

Recall that we used the following data from 3 World Cup tournaments (a total of 192 games) to introduce the idea of a **weighted average**.

We've added a third column to our table that gives us relative frequencies.

total # goals/game	frequency	relative frequency
0	17	17 / 192 = 0.089
1	45	45 / 192 = 0.234
2	51	51 / 192 = 0.266
3	37	37 / 192 = 0.193
4	25	25 / 192 = 0.130
5	11	11 / 192 = 0.057
6	3	3 / 192 = 0.016
7	2	2 / 192 = 0.010
8	1	1 / 192 = 0.005

The mean for this data is:

$$\frac{0(17)+1(45)+2(51)+3(37)+4(25)+5(11)+6(3)+7(2)+8(1)}{192}$$

Distributing the division by 192 we get:

$$0\left(\frac{17}{192}\right)+1\left(\frac{45}{192}\right)+2\left(\frac{51}{192}\right)+\cdots+8\left(\frac{1}{192}\right)$$

Notice that the mean is each number of goals per game multiplied by its relative frequency.

Since we usually write the relative frequencies as decimals, we can see that:

Mean number of goals per game =

- **0(0.089) + 1(0.234) + 2(0.266) + 3(0.193) + 4(0.130) + 5(0.057) + 6(0.016) + 7(0.010) + 8(0.005)**

= **2.36**, rounded to two decimal places.

In Exploratory Data Analysis, we used the **mean** of a sample of quantitative values—their arithmetic average—to tell the **center** of their distribution. We also saw how a weighted mean was used when we had a frequency table. These frequencies can be changed to relative frequencies.

So we are essentially using the relative frequency approach to find probabilities. We can use this to find the **mean**, or **center**, of a **probability distribution for a discrete random variable**, which will be a weighted average of its values; the more probable a value is the more weight it gets.

As always, it is important to distinguish between a concrete sample of observed values for a variable versus an abstract population of all values taken by a random variable in the long run.

Whereas we denoted the mean of a sample as x-bar, we now denote the mean of a random variable using the **Greek letter mu** with a subscript for the random variable we are using.

Let's see how this is done by looking at a specific example.

EXAMPLE: Xavier's Production Line

Xavier's production line produces a variable number of defective parts in an hour, with probabilities shown in this table:

X	0	1	2	3	4
P(X=x)	.15	.30	.25	.20	.10

How many defective parts are typically produced in an hour on Xavier's production line? If we sum up the possible values of X, each weighted with its probability, we have

$$\mu_X = 0(0.15)+1(0.30)+2(0.25)+3(0.20)+4(0.10)=1.8$$

Here is the general definition of the mean of a discrete random variable:

In general, for any discrete random variable X with probability distribution

X	x_1	x_2	x_3	⋯	x_n
P(X=x)	p_1	p_2	p_3	⋯	p_n

The **mean** of X is defined to be

$$\mu_X = x_1p_1 + x_2p_2 + \ldots + x_np_n = \sum_{i=1}^{n} x_ip_i$$

- In general, the mean of a random variable tells us its "long-run" average value.
- It is sometimes referred to as the **expected value**of the random variable.

Although "**expected value**" is a common, and even preferred term in the field of statistics, this expression may be somewhat misleading, because in many cases it is impossible for a random variable to actually equal its expected value.

For example, the mean number of goals for a World Cup soccer game is 2.36. But we can never expect any single game to result in 2.36 goals, since it is not possible to score a fraction of a goal. Rather, 2.36 is the long-run average of all World Cup soccer games.

In the case of Xavier's production line, the mean number of defective parts produced in an hour is 1.8. But the actual number of defective parts produced in any given hour can never equal 1.8, since it must take whole number values.

To get a better feel for the mean of a random variable, let's extend the defective parts example:

EXAMPLE: Xavier's and Yves' Production Lines

Recall the probability distribution of the random variable X, representing the number of defective parts in an hour produced by Xavier's production line.

X	0	1	2	3	4
P(X=x)	.15	.30	.25	.20	.10

The number of defective parts produced each hour by Yves' production line is a random variable Y with the following probability distribution:

Y	0	1	2	3	4
P(Y=y)	.05	.05	.10	.75	.05

Look at both probability distributions. Both X and Y take the same possible values (0, 1, 2, 3, 4).

However, they are very different in the way the probability is distributed among these values.

Learn by Doing: Comparing Probability Distributions #1

Did I Get This?: Mean of Discrete Random Variable

Variance and Standard Deviation of a Discrete Random Variable

Learning Objectives

LO 6.13: Find the mean, variance, and standard deviation of a discrete random variable.

In Exploratory Data Analysis, we used the mean of a sample of quantitative values (their arithmetic average, x-bar) to tell the center of their distribution, and the standard deviation (s) to tell the typical distance of sample values from their mean.

We described the center of a probability distribution for a random variable by reporting its mean which we denoted by the Greek letter mu.

Now we would like to establish an accompanying measure of **spread**.

Our measure of spread will still report the typical distance of values from their means, but in order to distinguish the spread of a population of all of a random variable's values from the spread (s) of sample values, we will denote the standard deviation of the random variable X with the Greek lower case "**sigma**," and use a subscript to remind us what is the variable of interest (there may be more than one in later problems):

We will also focus more frequently than before on the squared standard deviation, called the **variance**, because some important rules we need to invoke are in terms of variance rather than standard deviation.

EXAMPLE: Xavier's Production Line

Recall that the number of defective parts produced each hour by Xavier's production line is a random variable X with the following probability distribution:

X	0	1	2	3	4
P(X=x)	.15	.30	.25	.20	.10

We found the mean number of defective parts produced per hour to be 1.8.

Obviously, there is variation about this mean: some hours as few as 0 defective parts are produced, whereas in other hours as many as 4 are produced.

Typically, how far does the number of defective parts fall from the mean of 1.8?

As we did for the spread of sample values, we measure the spread of a random variable by calculating the square root of the average squared deviation from the mean.

Now "average" is a weighted average, where more probable values of the random variable are accordingly given more weight.

Let's begin with the variance, or average squared deviation from the mean, and then take its square root to find the standard deviation:

Values of X	0	1	2	3	4
Dev. from mean	(0-1.8)	(1-1.8)	(2-1.8)	(3-1.8)	(4-1.8)
Sq. deviations	$(0-1.8)^2$	$(1-1.8)^2$	$(2-1.8)^2$	$(3-1.8)^2$	$(4-1.8)^2$
Probabilities	.15	.30	.25	.20	.10

$$\begin{aligned}\text{Variance} &= \sigma_X^2 = (0-1.8)^2(0.15)+(1-1.8)^2(0.30)+(2-1.8)^2(0.25)\\ &\quad +(3-1.8)^2(0.20)+(4-1.8)^2(0.1)\\ &= 1.46\end{aligned}$$

$$\text{standard deviation} = \sigma_X = \sqrt{1.46} = 1.21$$

How do we interpret the standard deviation of X?

- Xavier's production line produces an average of 1.80 defective parts per hour.
- **The number of defective parts varies from hour to hour; typically (or, on average), it is about 1.21 away from the mean 1.80.**

Here is the formal definition:

In general, for any discrete random variable X with probability distribution

X	x_1	x_2	x_3	...	x_n
P(X=x)	p_1	p_2	p_3	...	p_n

The **variance** of X is defined to be

$$\begin{aligned}\sigma_X^2 &= (x_1-\mu_X)^2 p_1 + (x_2-\mu_X)^2 p_2 + \ldots + (x_n-\mu_X)^2 p_n\\ &= \sum_{i=1}^{n} (x_i-\mu_X)^2 p_i\end{aligned}$$

There is also a "short-cut" formula which is faster for by-hand calculation. In the formula below we have dropped the subscript for the variable in the notation.In this short-cut, we simply need to

- square each X,
- multiply by the probability of that X,
- then sum those values.
- From that result we subtract the square of the mean to find the variance.

$$\text{Var}(X) = \sigma^2 = \textstyle\sum_{i=1}^{n} \left[x_i^2 P\left(X=x_i\right)\right] - \mu^2$$

The **standard deviation** is the square root of the variance

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

Did I Get This?: Standard Deviation of a Discrete Random Variable

The purpose of the next activity is to give you better intuition about the mean and standard deviation of a random variable.

Learn by Doing: Comparing Probability Distributions #2

EXAMPLE: Xavier's and Yves' Production Lines

Recall the probability distribution of the random variable X, representing the number of defective parts per hour produced by Xavier's production line, and the probability distribution of the random variable Y, representing the number of defective parts per hour produced by Yves' production line:

X	0	1	2	3	4
P(X=x)	.15	.30	.25	.20	.10

Y	0	1	2	3	4
P(Y=y)	.05	.05	.10	.75	.05

Look carefully at both probability distributions. Both X and Y take the same possible values (0, 1, 2, 3, 4).

However, they are very different in the way the probability is distributed among these values. We saw before that this makes a difference in means:

$$\mu_X = 1.8$$

$$\mu_Y = 2.7$$

We now want to get a sense about how the different probability distributions impact their standard deviations.

Recall that the **standard deviation** of a **random variable** can be **interpreted** as a **typical** (or the **long-run average**) **distance** between the **value of X and its mean**.

Learn by Doing: Comparing Probability Distributions #3

So, 75% of the time Y will assume a value (3) that is very close to its mean (2.7), while X will assume a value (2) that is close to its mean (1.8) much less often—only 25% of the time.

The long-run average, then, of the distance between the values of Y and their mean will be much smaller than the long-run average of the distance between the values of X and their mean.

Therefore

$$\sigma_Y < \sigma_X = 1.21$$

Actually we have

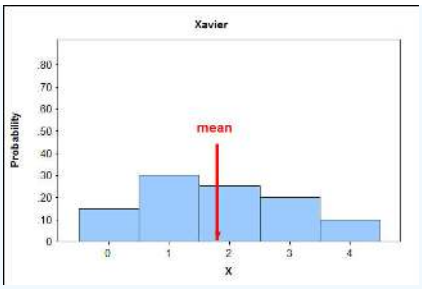
$$\sigma_Y = 0.85$$

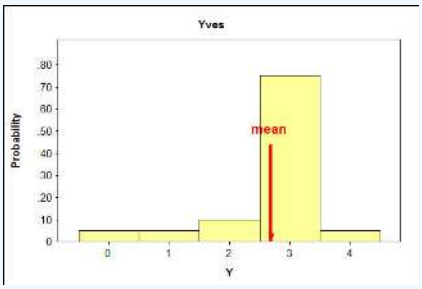
So we can draw the following conclusion:

Yves' production line produces an average of 2.70 defective parts per hour.

The number of defective parts varies from hour to hour; typically (or, on average), it is about 0.85 away from 2.70.

Here are the histograms for the production lines:





When we compare distributions, the distribution in which it is **more likely** to find values that are further from the mean will have a **larger** standard deviation.

Likewise, the distribution in which it is **less likely** to find values that are further from the mean will have the **smaller** standard deviation.

Did I Get This?: Standard Deviation of a Discrete Random Variable #2

Comment:

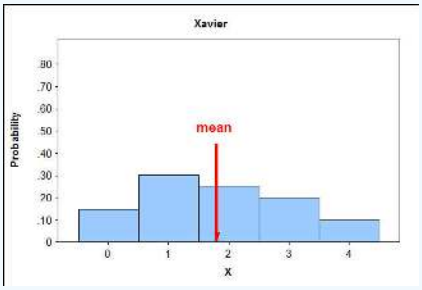
As we have stated before, using the mean and standard deviation gives us another way to assess which values of a random variable are unusual.

For reasonably symmetric distributions, any values of a random variable that fall within 2 or 3 standard deviations of the mean would be considered ordinary (not unusual).

For any distribution, it is unusual for values to fall outside of 3 or 4 standard deviations – depending on your definition of "unusual."

EXAMPLE: Xavier's Production Line-Unusual or Not?

Looking once again at the probability distribution for Xavier's production line:



Would it be considered unusual to have 4 defective parts per hour?

We know that the mean is 1.8 and the standard deviation is 1.21.

Ordinary values are within 2 (or 3) standard deviations of the mean.

- 1.8 – 2(1.21) = -0.62 and

- 1.8 + 2(1.21) = 4.22.

This gives us an interval from -0.62 to 4.22.

Since we cannot have a negative number of defective parts, the interval is essentially from 0 to 4.22.

Because 4 is within this interval, it would be considered ordinary. Therefore, it is **not unusual**.

Would it be considered unusual to have no defective parts?

Zero is within 2 standard deviations of the mean, so it would not be considered unusual to have no defective parts.

The following activity will reinforce this idea.

Learn by Doing: Unusual or Not?

Normal Random Variables

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

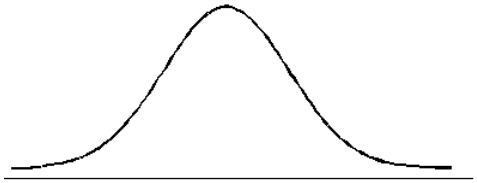
Learning Objectives

LO 6.2: Apply the standard deviation rule to the special case of distributions having the “normal” shape.

Video

Video: Normal Random Variables (2:08)

In the Exploratory Data Analysis unit of this course, we encountered data sets, **such as lengths of human pregnancies**, whose distributions naturally followed a symmetric unimodal bell shape, bulging in the middle and tapering off at the ends.

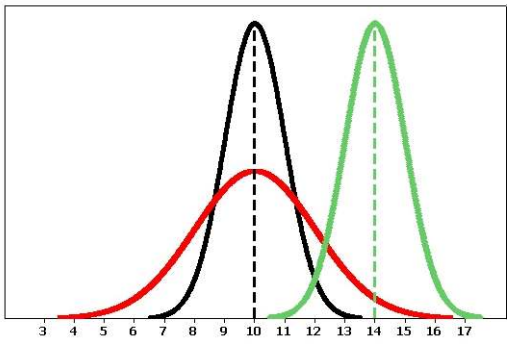


Many variables, such as pregnancy lengths, shoe sizes, foot lengths, and other human physical characteristics exhibit these properties: symmetry indicates that the variable is just as likely to take a value a certain distance below its mean as it is to take a value that same distance above its mean; the bell-shape indicates that values closer to the mean are more likely, and it becomes increasingly unlikely to take values far from the mean in either direction.

The particular shape exhibited by these variables has been studied since the early part of the nineteenth century, when they were first called “normal” as a way of suggesting their depiction of a common, natural pattern.

Observations of Normal Distributions

There are many normal distributions. Even though all of them have the bell-shape, they vary in their center and spread.



More specifically, the shape of the distribution is determined by its **mean** (mu, μ) and the spread is determined by its standard deviation (sigma, σ).

Some observations we can make as we look at this graph are:

- The black and the red normal curves have means or centers at μ = mu = 10. However, the red curve is more spread out and thus has a larger standard deviation. As you look at these two normal curves, notice that as the red graph is squished down, the spread gets larger, thus allowing the area under the curve to remain the same.
- The black and the green normal curves have the same standard deviation or spread (the range of the black curve is 6.5-13.5, and the green curve’s range is 10.5-17.5).

Even more important than the fact that many variables themselves follow the normal curve is the role played by the normal curve in sampling theory, as we'll see in the next section in our unit on probability.

Understanding the normal distribution is an important step in the direction of our overall goal, which is to relate sample means or proportions to population means or proportions. The goal of this section is to better understand normal random variables and their distributions.

The Standard Deviation Rule for Normal Random Variables

We began to get a feel for normal distributions in the Exploratory Data Analysis (EDA) section, when we introduced the Standard Deviation Rule (or the **68-95-99.7** rule) for how values in a normally-shaped **sample data set** behave relative to their sample mean (x-bar) and sample standard deviation (s).

This is the same rule that dictates how the distribution of a normal **random variable** behaves relative to its mean (mu, μ) and standard deviation (sigma, σ). Now we use probability language and notation to describe the random variable's behavior.

For example, in the EDA section, we would have said "68% of pregnancies in our data set fall within 1 standard deviation (s) of their mean (x-bar)." The analogous statement now would be "If X, the length of a randomly chosen pregnancy, is normal with mean (mu, μ) and standard deviation (sigma, σ), then

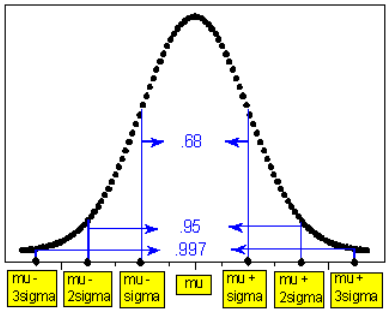
$$0.68 = P(\mu - \sigma < X < \mu + \sigma)$$

In general, if X is a normal random variable, then the probability is

- 68% that X falls within 1 standard deviation (sigma, σ) of the mean (mu, μ)
- 95% that X falls within 2 standard deviations (sigma, σ) of the mean (mu, μ)
- 99.7% that X falls within 3 standard deviation (sigma, σ) of the mean (mu, μ).

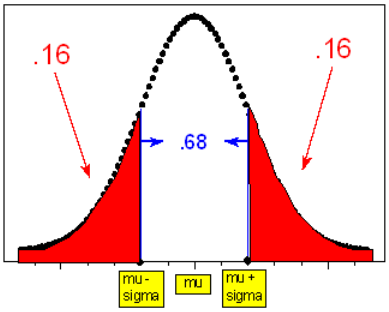
Using probability notation, we may write

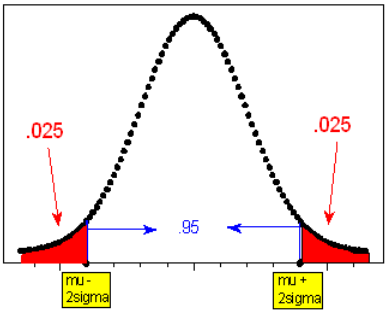
$$0.68 = P(\mu - \sigma < X < \mu + \sigma)$$
$$0.95 = P(\mu - 2\sigma < X < \mu + 2\sigma)$$
$$0.997 = P(\mu - 3\sigma < X < \mu + 3\sigma)$$

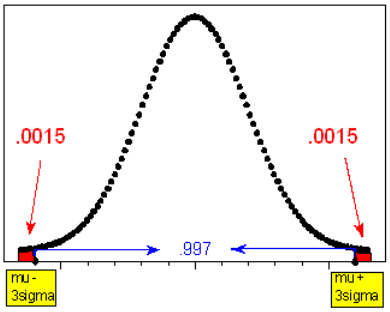


Comment

- Notice that the information from the rule can be interpreted from the perspective of the tails of the normal curve:
  - Since 0.68 is the probability of being within 1 standard deviation of the mean, (1 – 0.68) / 2 = 0.16 is the probability of being further than 1 standard deviation below the mean (or further than 1 standard deviation above the mean.)
  - Likewise, (1 – 0.95) / 2 = 0.025 is the probability of being more than 2 standard deviations below (or above) the mean.
  - And (1 – 0.997) / 2 = 0.0015 is the probability of being more than 3 standard deviations below (or above) the mean.
- The three figures below illustrate this.

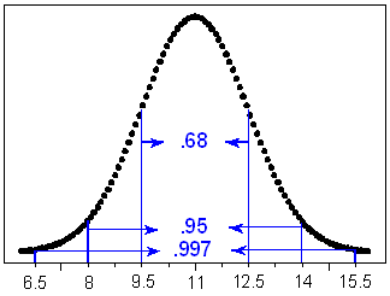






EXAMPLE: Foot Length

Suppose that foot length of a randomly chosen adult male is a normal random variable with mean μ = mu = 11 and standard deviation σ = sigma =1.5. Then the Standard Deviation Rule lets us sketch the probability distribution of X as follows:



(a) What is the probability that a randomly chosen adult male will have a foot length between 8 and 14 inches?

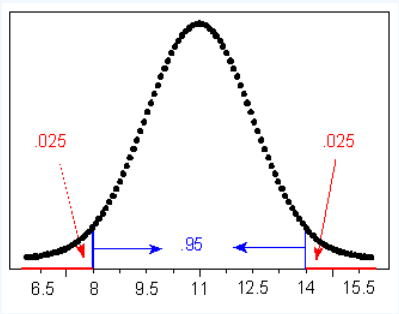
0.95, or 95%.

(b) An adult male is almost guaranteed (.997 probability) to have a foot length between what two values?

6.5 and 15.5 inches.

(c) The probability is only 2.5% that an adult male will have a foot length greater than how many inches?

14. (See image below)



Now you should try a few. (Use the figure that is just before **part (a)** to help you.)

Learn by Doing: Using the Standard Deviation Rule

Comment

- Notice that there are two types of problems we may want to solve: those like **(a)**, **(d)** and **(e)**, in which a particular interval of values of a normal random variable is given, and we are asked to find a probability, and those like **(b)**, **(c)** and **(f)**, in which a probability is given and we are asked to identify what the normal random variable's values would be.

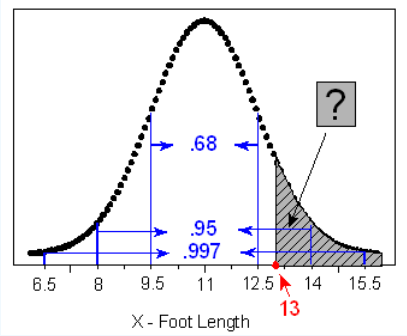
Did I Get This?: Using the Standard Deviation Rule

Learn by Doing: Normal Random Variables

Let's go back to our example of foot length:

EXAMPLE: Foot Length

How likely or unlikely is it for a male's foot length to be more than 13 inches?

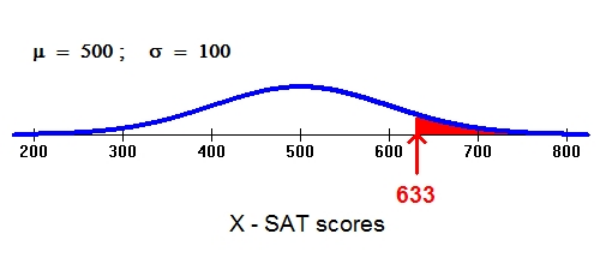


Since 13 inches doesn't happen to be exactly 1, 2, or 3 standard deviations away from the mean, we would only be able to give a very rough estimate of the probability at this point.

Clearly, the Standard Deviation Rule only describes the tip of the iceberg, and while it serves well as an introduction to the normal curve, and gives us a good sense of what would be considered likely and unlikely values, it is very limited in the probability questions it can help us answer.

Here is another familiar normal distribution:

EXAMPLE: SAT Scores



Suppose we are interested in knowing the probability that a randomly selected student will score 633 or more on the math portion of his or her SAT (this is represented by the red area). Again, 633 does not fall exactly 1, 2, or 3 standard deviations above the mean.

Notice, however, that an SAT score of 633 and a foot length of 13 are both about 1/3 of the way between 1 and 2 standard deviations. As you continue to read, you'll realize that this positioning relative to the mean is the key to finding probabilities.

Standard Normal Distribution

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Video: Standard Normal Distribution (4:12)

Finding Probabilities for a Normal Random Variable

Learning Objectives

LO 6.17: Find probabilities associated with a specified normal distribution.

As we saw, the Standard Deviation Rule is very limited in helping us answer probability questions, and basically limited to questions involving values that fall exactly 1, 2, and 3 standard deviations away from the mean. How do we answer probability questions in general? The key is the position of the value relative to the mean, measured in standard deviations.

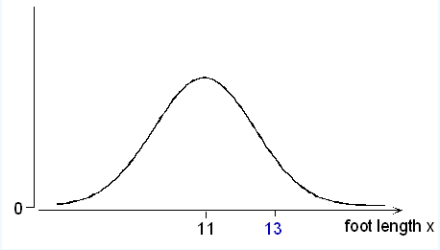
We can approach the answering of probability questions two possible ways: a table and technology. In the next sections, you will learn how to use the "standard normal table," and then how the same calculations can be done with technology.

Standardizing Values

The first step to assessing a probability associated with a normal value is to determine the **relative** value with respect to all the other values taken by that normal variable. This is accomplished by determining how many standard deviations below or above the mean that value is.

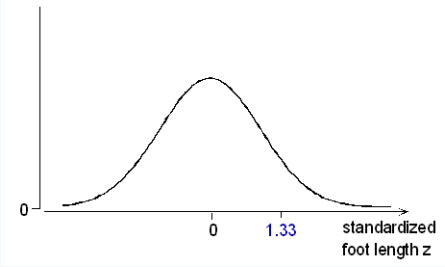
EXAMPLE: Foot Length

How many standard deviations below or above the mean male foot length is 13 inches? Since the mean is 11 inches, 13 inches is 2 inches above the mean.



Since a standard deviation is 1.5 inches, this would be 2 / 1.5 = 1.33 standard deviations above the mean. Combining these two steps, we could write:

(13 in. – 11 in.) / (1.5 inches per standard deviation) = (13 – 11) / 1.5 standard deviations = +1.33 standard deviations.



In the language of statistics, we have just found the **z-score** for a male foot length of 13 inches to be z = +1.33. Or, to put it another way, we have **standardized** the value of 13.

In general, the **standardized value** z tells how many standard deviations below or above the mean the original value is, and is calculated as follows:

z-score = (value – mean)/standard deviation

The convention is to denote a value of our normal random variable X with the letter "x."

$$z = \frac{x - \mu}{\sigma}$$

Notice that since the standard deviation (sigma, σ) is always positive, for values of x above the mean (mu, μ), z will be positive; for values of x below the mean (mu, μ), z will be negative.

Let's go back to our foot length example, and answer some more questions.

EXAMPLE: Foot Length

(a) What is the standardized value for a male foot length of 8.5 inches? How does this foot length relate to the mean?

z = (8.5 – 11) / 1.5 = -1.67. This foot length is 1.67 standard deviations **below** the mean.

(b) A man's standardized foot length is +2.5. What is his actual foot length in inches?

If z = +2.5, then his foot length is 2.5 standard deviations above the mean. Since the mean is 11, and each standard deviation is 1.5, we get that the man's foot length is: 11 + 2.5(1.5) = 14.75 inches.

Note that z-scores also allow us to compare values of different normal random variables. Here is an example:

(c) In general, women's foot length is shorter than men's. Assume that women's foot length follows a normal distribution with a mean of 9.5 inches and standard deviation of 1.2. Ross' foot length is 13.25 inches, and Candace's foot length is only 11.6 inches. Which of the two has a longer foot relative to his or her gender group?

To answer this question, let's find the z-score of each of these two normal values, bearing in mind that each of the values comes from a different normal distribution.

Ross: z-score = (13.25 – 11) / 1.5 = 1.5 (Ross' foot length is 1.5 standard deviations above the mean foot length for men).

Candace: z-score = (11.6 – 9.5) / 1.2 = 1.75 (Candace's foot length is 1.75 standard deviations above the mean foot length for women).

Note that even though Ross' foot is longer than Candace's, Candace's foot is longer relative to their respective genders.

Comment:

- Part (c) above illustrates how z-scores become crucial when you want to **compare distributions**.

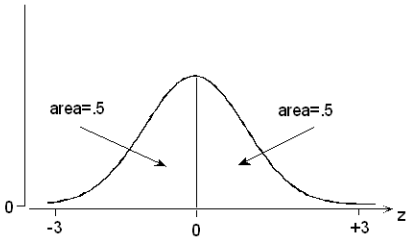
Did I Get This?: Standardized Scores (z-scores)

Finding Probabilities with the Normal Calculator and Table

Now that you have learned to assess the relative value of any normal value by standardizing, the next step is to evaluate probabilities. In other contexts, as mentioned before, we will first take the conventional approach of referring to a **normal table**, which tells the probability of a normal variable taking a value **less than** any standardized score z.

Standard Normal Table

Since normal curves are symmetric about their mean, it follows that the curve of z scores must be symmetric about 0. Since the total area under any normal curve is 1, it follows that the areas on either side of z = 0 are both 0.5. Also, according to the Standard Deviation Rule, most of the area under the standardized curve falls between z = -3 and z = +3.



The normal table outlines the precise behavior of the standard normal random variable Z, the number of standard deviations a normal value x is below or above its mean. The normal table provides probabilities that a standardized normal random variable Z would take a value less than or equal to a particular value z*.

These particular values are listed in the form *.* in rows along the left margins of the table, specifying the ones and tenths. The columns fine-tune these values to hundredths, allowing us to look up the probability of being below any standardized value z of the form *.**.

For example, in the part of the table shown below, we can see that for a z-score of -2.81, we would find P(Z < -2.81) = 0.0025.

Standard normal probabilities

z	.00	.01	.02
−3.4	.0003	.0003	.0003
−3.3	.0005	.0005	.0005
−3.2	.0007	.0007	.0006
−3.1	.0010	.0009	.0009
−3.0	.0013	.0013	.0013
−2.9	.0019	.0018	.0018
−2.8	.0026	.0025	.0024
−2.7	.0035	.0034	.0033
−2.6	.0047	.0045	.0044

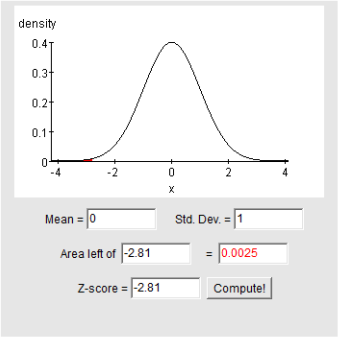
By construction, the probability P(Z < z*) equals the area under the z curve to the left of that particular value z*.

A quick sketch is often the key to solving normal problems easily and correctly.

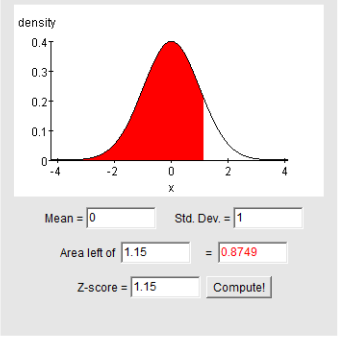
Although normal tables are the traditional way to solve these problems, you can also use the normal calculator.

Normal Distribution Calculator: Non-JAVA Version

The image below illustrates the results of using the online calculator to find P(Z < -2.81) and P(Z < 1.15). Notice that the calculator behaves exactly as the table.



P(Z < -2.81) = 0.0025



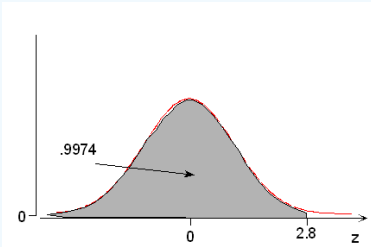
P(Z < 1.15) = 0.8749

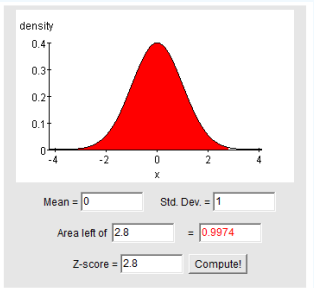
It is your choice to use the table or the online calculator but we will usually illustrate with the online calculator.

EXAMPLE: Standard Normal Probabilities

(a) What is the probability of a normal random variable taking a value less than 2.8 standard deviations above its mean?

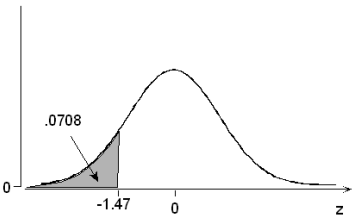
P(Z < 2.8) = 0.9974 or 99.74%.

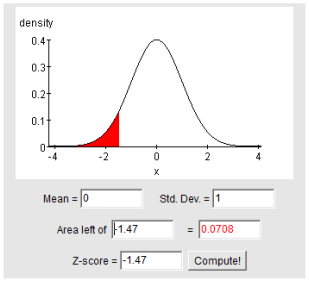




(b) What is the probability of a normal random variable taking a value lower than 1.47 standard deviations below its mean?

P(Z < -1.47) = 0.0708, or 7.08%.





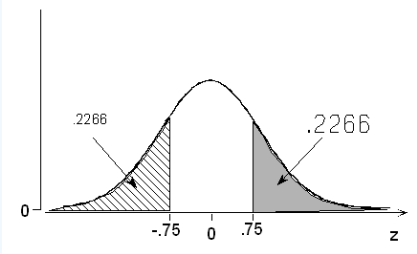
(c) What is the probability of a normal random variable taking a value **more** than 0.75 standard deviations above its mean?

The fact that the problem involves the word "more" rather than "less" should not be overlooked! Our normal calculator provides left-tail probabilities, and adjustments must be made for any other type of problem.

Method 1:

By symmetry of the z curve centered on 0,

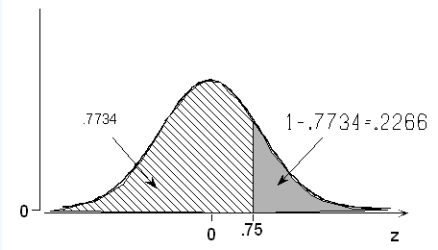
P(Z > +0.75) = P(Z < -0.75) = 0.2266.



Method 2:

Because the total area under the normal curve is 1,

P(Z > +0.75) = 1 – P(Z < +0.75) = 1 – 0.7734 = 0.2266.

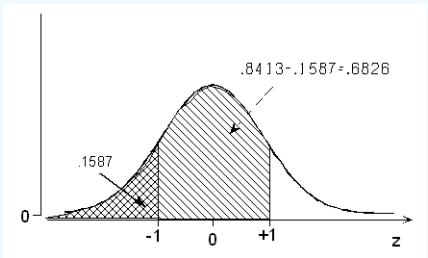


[**Note:** most students prefer to use Method 1, which does not require subtracting 4-digit probabilities from 1.]

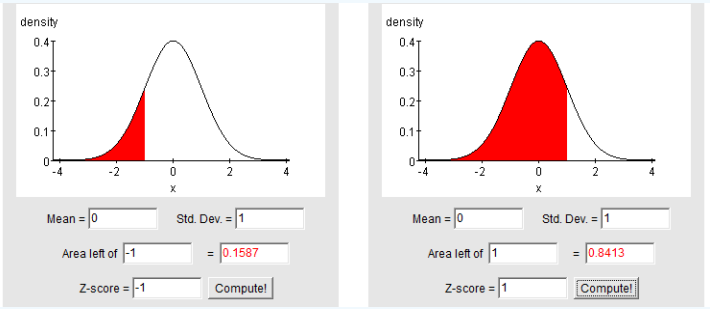
(d) What is the probability of a normal random variable taking a value between 1 standard deviation below and 1 standard deviation above its mean?

To find probabilities in between two standard deviations, we must put them in terms of the probabilities below. A sketch is especially helpful here:

P(-1 < Z < +1) = P(Z < +1) – P(Z < -1) = 0.8413 – 0.1587 = 0.6826.



Here are the normal calculator results which would be needed.



Did I Get This?: Standard Normal Probabilities

Comments:

- So far, we have used the normal calculator or table to find a probability, given the number (z) of standard deviations below or above the mean. The solution process when using the table involved first locating the given z value of the form *.** in the margins, then finding the corresponding probability of the form 0.**** inside the table as our answer.
- Now, in Example 2, a probability will be given and we will be asked to find a z value. The solution process using the table involves first locating the given probability of the form 0.**** inside the table, then finding the corresponding z value of the form *.** as our answer. For the online calculator, the solution is as simply typing in the correct probability and having the calculator solve, in reverse, for the z-score.

Finding Standard Normal Scores

Learning Objectives

LO 6.18: Given a probability, find scores associated with a specified normal distribution.

It is often good to think about this process as the reverse of finding probabilities. In these problems, we will be given some information about the area in a range and asked to provide the z-score(s) associated with that range. Common types of questions are

- Find the standard normal z-score corresponding to the top (or bottom) 8%.
- Find the standard normal z-score associated with the 25th percentile.
- Find the standard normal z-scores which contain the middle 40%.

EXAMPLE: Given Probabilities - Find Z-Scores

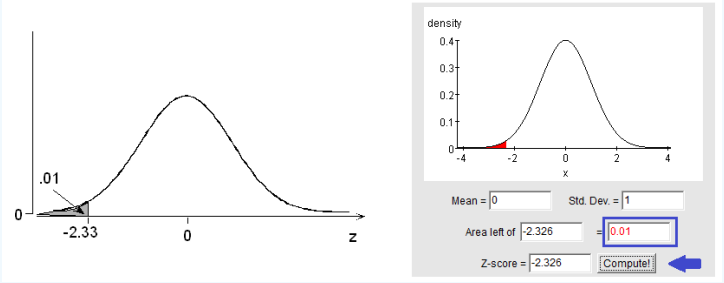
(a) What standard normal z-score is associated with the bottom (or lowest) 1%? The probability is 0.01 that a standardized normal variable takes a value **below** what particular value of z?

The closest we can come to a probability of 0.01 inside the table is 0.0099, in the z = -2.3 row and 0.03 column: **z = -2.33**. In other words, the probability is 0.01 that the value of a normal variable is lower than 2.33 standard deviations below its mean.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

Using the online calculator, we simply use the calculator in reverse by typing in 0.01 in the “area” box (outlined in blue) and then click “compute” to see the associated z-score. Remember that, like the table, we always need to provide this

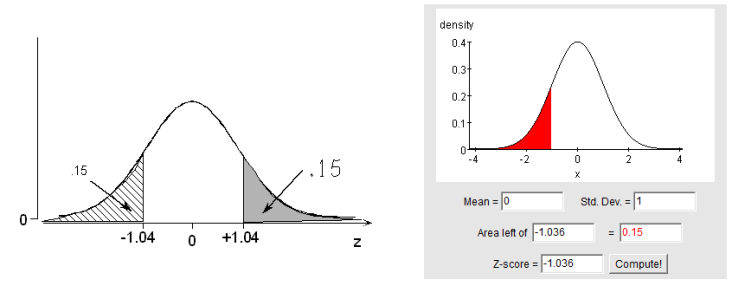
calculator with the **area to the left** of the z-score we are currently trying to find.



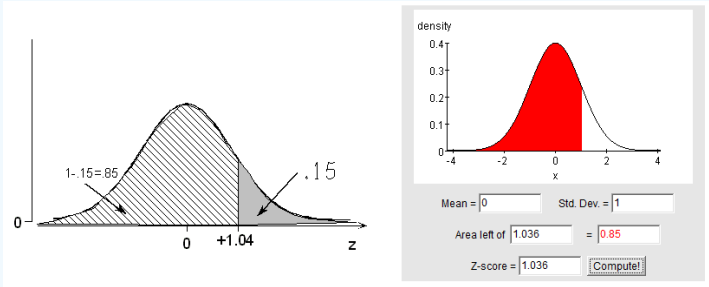
(b) What standard normal z-score corresponds to the top (or upper) 15%? The probability is 0.15 that a standardized normal variable takes a value **above** what particular value of z?

Remember that the calculator and table only provide probabilities of being **below** a certain value, not above. Once again, we must rely on one of the properties of the normal curve to make an adjustment.

Method 1: According to the table, 0.15 (actually 0.1492) is the probability of being **below** -1.04. By symmetry, 0.15 must also be the probability of being **above** +1.04. Using the calculator, we can enter 0.15 exactly and find that the corresponding z-score is actually -1.036 giving a final answer of z = +1.036 or +1.04 if we round to two decimal places which is our preference (this results in no differences for students who use the table or the online calculator).



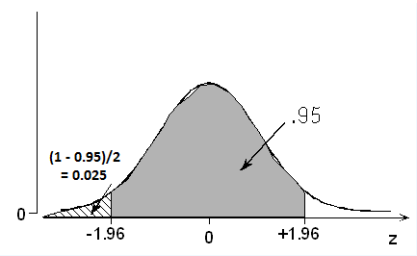
Method 2: If 0.15 is the probability of being above the value we seek, then 1 – 0.15 = 0.85 must be the probability of being below the value we seek. According to the table, 0.85 (actually 0.8508) is the probability of being below +1.04.

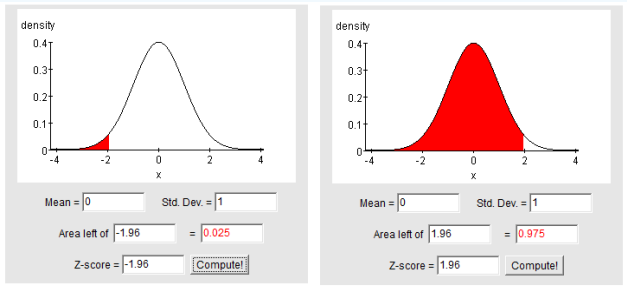


In other words, we have found 0.15 to be the probability that a normal variable takes a value more than 1.04 standard deviations above its mean.

(c) What standard normal z-scores contain the middle 95%? The probability is 0.95 that a normal variable takes a value within how many standard deviations of its mean?

A symmetric area of 0.95 centered at 0 extends to values -z* and +z* such that the remaining (1 – 0.95) / 2 = 0.025 is below -z* and also 0.025 above +z*. The probability is 0.025 that a standardized normal variable is below -1.96. Thus, the probability is 0.95 that a normal variable takes a value within 1.96 standard deviations of its mean. Once again, the Standard Deviation Rule is shown to be just roughly accurate, since it states that the probability is 0.95 that a normal variable takes a value within 2 standard deviations of its mean.





Did I Get This?: Finding Standard Normal Scores

Although the online calculator can provide results for any probability or z-score, our standard normal table, like most, only provides probabilities for z values between -3.49 and +3.49. The following example demonstrates how to handle cases where z exceeds 3.49 in absolute value.

EXAMPLE: Extreme Probabilities

(a) What is the probability of a normal variable being lower than 5.2 standard deviations below its mean?

There is no need to panic about going "off the edge" of the normal table. We already know from the Standard Deviation Rule that the probability is only about (1 -0 .997) / 2 = 0.0015 that a normal value would be more than 3 standard deviations away from its mean in one direction or the other. The table provides information for z values as extreme as plus or minus 3.49: the probability is only 0.0002 that a normal variable would be lower than 3.49 standard deviations below its mean. Any more standard deviations than that, and we generally say the probability is approximately zero.

In this case, we would say the probability of being lower than 5.2 standard deviations below the mean is approximately zero:

P(Z < -5.2) = 0 (approx.)

(b) What is the probability of the value of a normal variable being higher than 6 standard deviations below its mean?

Since the probability of being lower than 6 standard deviations below the mean is approximately zero, the probability of being higher than 6 standard deviations below the mean must be approximately 1.

P(Z > -6) = 1 (approx.)

(c) What is the probability of a normal variable being less than 8 standard deviations above the mean?

Approximately 1. P(Z < +8) = 1 (approx.)

(d) What is the probability of a normal variable being greater than 3.5 standard deviations above the mean?

Approximately 0. P(Z > +3.5) = 0 (approx.)

Normal Applications

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Video: Normal Applications (9:41)

Working with Non-standard Normal Values

Learning Objectives

LO 6.17: Find probabilities associated with a specified normal distribution.

In a much earlier example, we wondered,

"How likely or unlikely is a male foot length of more than 13 inches?" We were unable to solve the problem, because 13 inches didn't happen to be one of the values featured in the Standard Deviation Rule.

Subsequently, we learned how to standardize a normal value (tell how many standard deviations below or above the mean it is) and how to use the normal calculator or table to find the probability of falling in an interval a certain number of standard deviations below or above the mean.

By combining these two skills, we will now be able to answer questions like the one above.

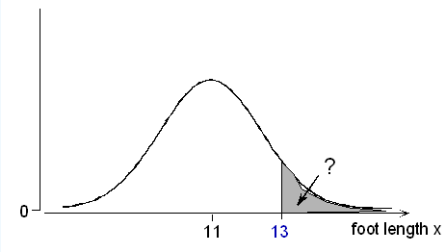
To convert between a non-standard normal (X) and the standard normal (Z) use the following equations, as needed:

$$Z = \frac{x - \mu}{\sigma} \qquad X = \mu + zQ$$

EXAMPLE: Male Foot Length

Male foot lengths have a normal distribution, with mean (mu, μ) = 11 inches, and standard deviation (sigma, σ) = 1.5 inches.

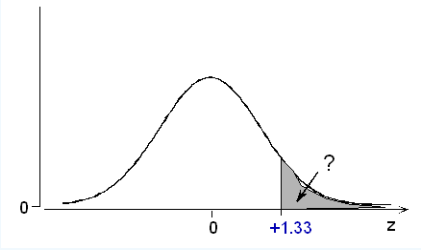
(a) What is the probability of a foot length of more than 13 inches?



First, we standardize:

$$z = \frac{x - \mu}{\sigma} = \frac{13 - 11}{1.5} = +1.33$$

The probability that we seek, P(X > 13), is the same as the probability that a normal variable takes a value greater than 1.33 standard deviations above its mean, i.e. P(Z > +1.33)



This can be solved with the normal calculator or table, after applying the property of symmetry:

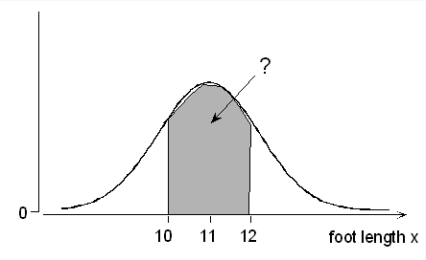
P(Z > +1.33) = P(Z < -1.33) = 0.0918.

A male foot length of more than 13 inches is on the long side, but not too unusual: its probability is about 9%.

We can streamline the solution in terms of probability notation and write:

P(X > 13) = P(Z > 1.33) = P(Z < −1.33) = 0.0918

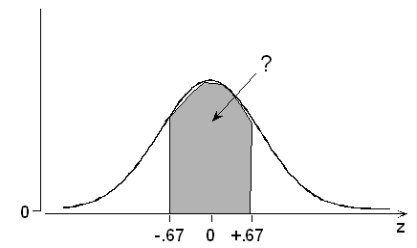
(b) What is the probability of a male foot length between 10 and 12 inches?



The standardized values of 10 and 12 are, respectively,

$$\frac{10-11}{1.5} = -0.67 \text{ and } \frac{12-11}{1.5} = 0.67$$

Note: The two z-scores in a "between" problem will not always be the same value. You must calculate both or, in this case, you could recognize that both values are the same distance from the mean and hence result in z-scores which are equal but of opposite signs.



P(-0.67 < Z < +0.67) = P(Z < +0.67) – P(Z < -0.67) = 0.7486 – 0.2514 = 0.4972.

Or, if you prefer the streamlined notation,

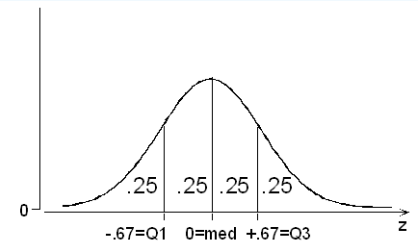
P(10 < X < 12) = P(−0.67 < Z < +0.67) = P(Z < +0.67) − P(Z < −0.67) = 0.7486 − 0.2514 = 0.4972.

Comments:

By solving the above example, we inadvertently discovered the quartiles of a normal distribution! P(Z < -0.67) = 0.2514 tells us that roughly 25%, or one quarter, of a normal variable's values are less than 0.67 standard deviations below the mean.

P(Z < +0.67) = 0.7486 tells us that roughly 75%, or three quarters, are less than 0.67 standard deviations above the mean.

And of course the median is equal to the mean, since the distribution is symmetric, the median is 0 standard deviations away from the mean.



Be sure to verify these results for yourself using the calculator or table!

Let's look at another example.

EXAMPLE: Length of a Human Pregnancy

Length (in days) of a randomly chosen human pregnancy is a normal random variable with mean (mu, μ) = 266 and standard deviation (sigma, σ) = 16.

(a) Find Q1, the median, and Q3. Using the z-scores we found in the previous example we have

Q1 = 266 – 0.67(16) = 255

median = mean = 266

Q3 = 266 + 0.67(16) = 277

Thus, the probability is 1/4 that a pregnancy will last less than 255 days; 1/2 that it will last less than 266 days; 3/4 that it will last less than 277 days.

(b) What is the probability that a randomly chosen pregnancy will last less than 246 days?

Since (246 – 266) / 16 = -1.25, we write

P(X < 246) = P(Z < −1.25) = 0.1056

(c) What is the probability that a randomly chosen pregnancy will last longer than 240 days?

Since (240 – 266) / 16 = -1.63, we write

P(X > 240) = P(Z > −1.63) = P(Z < +1.63) = 0.9484

Since the mean is 266 and the standard deviation is 16, most pregnancies last longer than 240 days.

(d) What is the probability that a randomly chosen pregnancy will last longer than 500 days?

Method 1:

Common sense tells us that this would be **impossible**.

Method 2:

The standardized value of 500 is (500 – 266) / 16 = +14.625.

P(X > 500) = P(Z > 14.625) = 0.

(e) Suppose a pregnant woman's husband has scheduled his business trips so that he will be in town between the 235th and 295th days. What is the probability that the birth will take place during that time?

The standardized values are (235 – 266) / 16) = -1.94 and (295 – 266) / 16 = +1.81.

P(235 < X < 295) = P(−1.94 < Z < +1.81) = P(Z < +1.81) − P(Z < −1.94) = 0.9649 − 0.0262 = 0.9387.

There is close to a 94% chance that the husband will be in town for the birth.

Be sure to verify these results for yourself using the calculator or table!

The purpose of the next activity is to give you guided practice at solving word problems that involve normal random variables. In particular, we'll solve problems like the examples you just went over, in which you are asked to find the probability that a normal random variable falls within a certain interval.

Learn by Doing: Find Normal Probabilities

The previous examples most followed the same general form: given values of a normal random variable, you were asked to find an associated probability. The two basic steps in the solution process were to

- **Standardize to Z;**
- **Find associated probabilities using the standard normal calculator or table.**

Finding Normal Scores

Learning Objectives

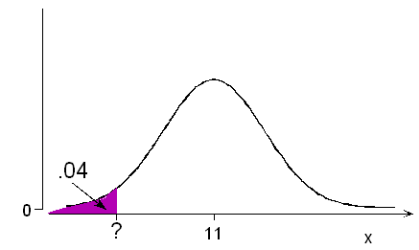
LO 6.18: Given a probability, find scores associated with a specified normal distribution.

The next example will be a different type of problem: given a certain probability, you will be asked to find the associated value of the normal random variable. The solution process will go more or less in reverse order from what it was in the previous examples.

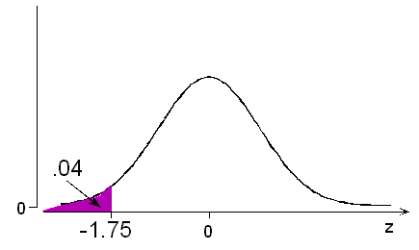
EXAMPLE: Foot Length

Again, foot length of a randomly chosen adult male is a normal random variable with a mean of 11 and standard deviation of 1.5.

(a) The probability is 0.04 that a randomly chosen adult male foot length will be less than how many inches?

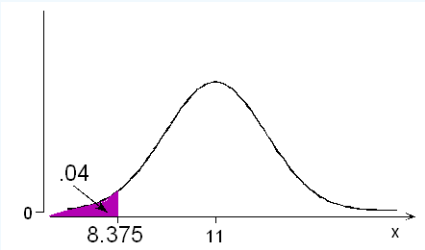


According to the normal calculator or table, a probability of 0.04 below (actually 0.0401) is associated with z = -1.75.



In other words, the probability is 0.04 that a normal variable takes a value lower than 1.75 standard deviations below its mean.

For adult male foot lengths, this would be 11 – 1.75(1.5) = 8.375. The probability is 0.04 that an adult male foot length would be less than 8.375 inches.



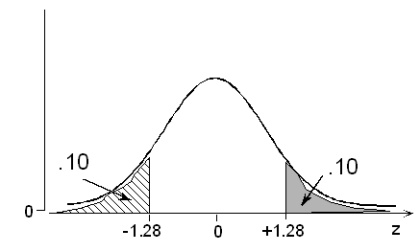
(b) The probability is 0.10 that an adult male foot will be longer than how many inches? Caution is needed here because of the word "longer."

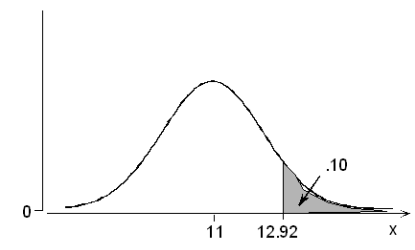
Once again, we must remind ourselves that the calculator and table only show the probability of a normal variable taking a value **lower than** a certain number of standard deviations below or above its mean. Adjustments must be made for problems that involve probabilities besides "lower than" or "less than." As usual, we have a choice of invoking either symmetry or the fact that the total area under the normal curve is 1. Students should examine both methods and decide which they prefer to use for their own purposes.

Method 1:

According to the calculator or table, a probability of 0.10 **below** is associated with a z value of -1.28. By symmetry, it follows that a probability of 0.10 **above** has z = +1.28.

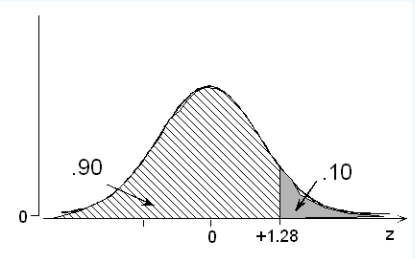
We seek the foot length that is 1.28 standard deviations above its mean: 11 + 1.28(1.5) = 12.92, or just under 13 inches.





Method 2: If the probability is 0.10 that a foot will be longer than the value we seek, then the probability is 0.90 that a foot will be shorter than that same value, since the probabilities must sum to 1.

According to the calculator or table, a probability of 0.90 below is associated with a z value of +1.28. Again, we seek the foot length that is 1.28 standard deviations above its mean, or 12.92 inches.



Comment:

- **Part (a) in the above example** could have been re-phrased as: "0.04 is the **proportion** of all adult male foot lengths that are below what value?", which takes the perspective of thinking about the probability as a proportion of occurrences in the long-run. As originally stated, it focuses on the chance of a randomly chosen individual having a normal value in a given interval.

EXAMPLE: Money Spent for Lunch

A study reported that the amount of money spent each week for lunch by a worker in a particular city is a normal random variable with a mean of $35 and a standard deviation of $5.

(a) The probability is 0.97 that a worker will spend less than how much money in a week on lunch?

The z associated with a probability of 0.9700 below is +1.88. The amount that is 1.88 standard deviations above the mean is **35 + 1.88(5) = 44.4, or $44.40.**

(b) There is a 30% chance of spending more than how much for lunches in a week?

The z associated with a probability of 0.30 above is +0.52. The amount is **35 + 0.52(5) = 37.6, or $37.60.**

Comment:

- Another way of expressing **Example (part a.)** above would be to ask, "What is the 97th **percentile** for the amount (X) spent by workers in a week for their lunch?" Many normal variables, such as heights, weights, or exam scores, are often expressed in terms of percentiles.

EXAMPLE:

The height X (in inches) of a randomly chosen woman is a normal random variable with a mean of 65 and a standard deviation of 2.5.

What is the height of a woman who is in the 80th percentile?

A probability of 0.7995 in the table corresponds to z = +0.84. Her height is **65 + 0.84(2.5) = 67.1 inches.**

By now we have had practice in solving normal probability problems in both directions: those where a normal value is given and we are asked to report a probability and those where a probability is given and we are asked to report a normal value. Strategies for solving such problems are outlined below:

- Given a normal value x, solve for probability:
  - Standardize: calculate

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

- If you are using the online calculator: Type the z-score for which you wish to find the area to the left and hit "compute."
- If you are using the table: Locate z in the margins of the normal table (ones and tenths for the row, hundredths for the column). Find the corresponding probability (given to four decimal places) of a normal random variable taking a value below z inside the table.
- (Adjust if the problem involves something other than a "less-than" probability, by invoking either symmetry or the fact that the total area under the normal curve is 1.)

- Given a probability, solve for normal value x:
  - (Adjust if the problem involves something other than a "less-than" probability, by invoking either symmetry or the fact that the total area under the normal curve is 1.)
  - Locate the probability (given to four decimal places) inside the normal table. Using the table, find the corresponding z value in the margins (row for ones and tenths, column for hundredths). Using the calculator, provide the area to left of the z-score you wish to find and hit "compute."
  - "Unstandardize": calculate

$$X = \mu + z\sigma$$

This next activity is a continuation of the previous one, and will give you guided practice in solving word problems involving the normal distribution. In particular, we'll solve problems like the ones you just solved, in which you are given a probability and you are asked to find the normal value associated with it.

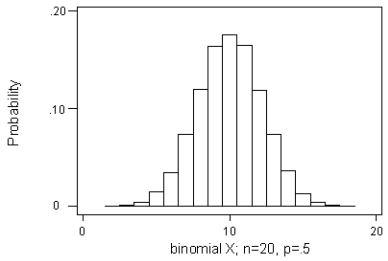
Learn by Doing: Find Normal Scores

Normal Approximation for Binomial

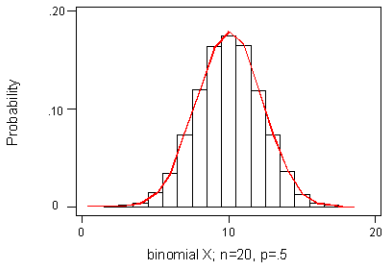
The normal distribution can be used as a reasonable approximation to other distributions under certain circumstances. Here we will illustrate this approximation for the binomial distribution.

We will not do any calculations here as we simply wish to illustrate the concept. In the next section on sampling distributions, we will look at another measure related to the binomial distribution, the sample proportion, and at that time we will discuss the underlying normal distribution.

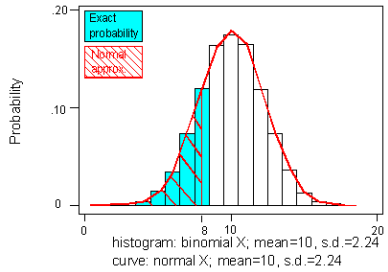
Consider the binomial probability distribution displayed below for n = 20 and p = 0.5.



Now we overlay a normal distribution with the same mean and standard deviation.



And in the final image, we can see the regions for the exact and approximate probabilities shaded.



Unfortunately, the approximated probability, 0.1867, is quite a bit different from the actual probability, 0.2517. However, this example constitutes something of a “worst-case scenario” according to the usual criteria for use of a normal approximation.

Rule of Thumb

Probabilities for a binomial random variable X with n and p may be approximated by those for a normal random variable having the same mean and standard deviation as long as the sample size n is large enough relative to the proportions of successes and failures, p and 1 – p. Our Rule of Thumb will be to require that

$np \geq 10$ and $n(1 - p) \geq 10$

Continuity Correction

It is possible to improve the normal approximation to the binomial by adjusting for the discrepancy that arises when we make the shift from the areas of histogram rectangles to the area under a smooth curve. For example, if we want to find the binomial probability that X is less than **or equal to** 8, we are including the area of the entire rectangle over 8, which actually extends to 8.5. Our normal approximation only included the area up to 8.

Summary (Unit 3B - Random Variables)

We have almost reached the end our discussion of probability. We were introduced to the important concept of **random variables**, which are quantitative variables whose value is determined by the outcome of a random experiment.

We discussed discrete and continuous random variables.

We saw that all the information about a **discrete random variable** is packed into its probability distribution. Using that, we can answer probability questions about the random variable and find its **mean and standard deviation**. We ended the part on discrete random variables by presenting a special class of discrete random variables – **binomial random variables.**

As we dove into **continuous random variables**, we saw how calculations can get complicated very quickly, when probabilities associated with a continuous random variable are found by calculating **areas under its density curve**.

As an example for a continuous random variable, we presented the **normal random variable**, and discussed it at length. The normal distribution is extremely important, not just because many variables in real life follow the normal distribution, but mainly because of the important role it plays in statistical inference, our ultimate goal of this course.

We learned how we can avoid calculus by using the **standard normal calculator or table** to find probabilities associated with the normal distribution, and learned how it can be used as an **approximation to the binomial** distribution under certain conditions.

Random Variables

A random variable is a variable whose values are numerical results of a random experiment.

- A **discrete random variable** is summarized by its probability distribution — a list of its possible values and their corresponding probabilities.

The sum of the probabilities of all possible values must be 1.

The probability distribution can be represented by a table, histogram, or sometimes a formula.

- The **probability distribution** of a random variable can be supplemented with numerical measures of the center and spread of the random variable.

Center: The center of a random variable is measured by its mean (which is sometimes also referred to as the **expected value**).

The mean of a random variable can be interpreted as its long run average.

The mean is a weighted average of the possible values of the random variable weighted by their corresponding probabilities.

Spread: The spread of a random variable is measured by its variance, or more typically by its standard deviation (the square root of the variance).

The standard deviation of a random variable can be interpreted as the typical (or long-run average) distance between the value that the random variable assumes and the mean of X.

Binomial Random Variables

- The binomial random variable is a type of discrete random variable that is quite common.
- The binomial random variable is defined in a random experiment that consists of n independent trials, each having two possible outcomes (called “success” and “failure”), and each having the same probability of success: p. Such a random experiment is called the binomial random experiment.
- The binomial random variable represents the number of successes (out of n) in a binomial experiment. It can therefore have values as low as 0 (if none of the n trials was a success) and as high as n (if all n trials were successes).
- There are “many” binomial random variables, depending on the number of trials (n) and the probability of success (p).
- The probability distribution of the binomial random variable is given in the form of a formula and can be used to find probabilities. Technology can be used as well.
- The mean and standard deviation of a binomial random variable can be easily found using short-cut formulas.

Continuous Random Variables

The probability distribution of a continuous random variable is represented by a probability density curve. The probability that the random variable takes a value in any interval of interest is the area above this interval and below the density curve.

An important example of a continuous random variable is the **normal random variable**, whose probability density curve is symmetric (bell-shaped), bulging in the middle and tapering at the ends.

- There are “many” normal random variables, each determined by its mean μ (mu) (which determines where the density curve is centered) and standard deviation σ (sigma) (which determines how spread out (wide) the normal density curve is).
- Any normal random variable follows the Standard Deviation Rule, which can help us find probabilities associated with the normal random variable.
- Another way to find probabilities associated with the normal random variable is using the standard normal table. This process involves finding the z-score of values, which tells us how many standard deviations below or above the mean the value is.
- An important application of the normal random variable is that it can be used as an approximation of the binomial random variable (under certain conditions). A continuity correction can improve this approximation.

CHAPTER OVERVIEW

Unit 3B: Sampling Distribution

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

NOTE: The following videos discuss all three pages related to sampling distributions.

Review: We will apply the concepts of normal random variables to **two random variables which are summary statistics from a sample** – these are the **sample mean (x-bar)** and the **sample proportion (p-hat)**.

- Normal Random Variables
  - Standard Normal Distribution
  - Normal Applications

Video

Video: Sampling Distributions (34:00 total time)

Introduction

Already on several occasions we have pointed out the important distinction between a **population** and a **sample**. In Exploratory Data Analysis, we learned to summarize and display values of a variable for a **sample**, such as displaying the blood types of 100 randomly chosen U.S. adults using a pie chart, or displaying the heights of 150 males using a histogram and supplementing it with appropriate numerical measures such as the sample mean (x-bar) and sample standard deviation (s).

In our study of Probability and Random Variables, we discussed the long-run behavior of a variable, considering the **population** of all possible values taken by that variable. For example, we talked about the distribution of blood types among all U.S. adults and the distribution of the random variable X, representing a male's height.

Now we focus directly on the relationship between the values of a variable for a **sample** and its values for the entire **population** from which the sample was taken. This material is the bridge between probability and our ultimate goal of the course, statistical inference. In inference, we look at a sample and ask what we can say about the population from which it was drawn.

Now, we'll pose the reverse question: **If I know what the population looks like, what can I expect the sample to look like?** Clearly, inference poses the more practical question, since in practice we can look at a sample, but rarely do we know what the whole population looks like. This material will be more theoretical in nature, since it poses a problem which is not really practical, but will present important ideas which are the foundation for statistical inference.

Parameters vs. Statistics

Learning Objectives

LO 6.19: Identify and distinguish between a parameter and a statistic.

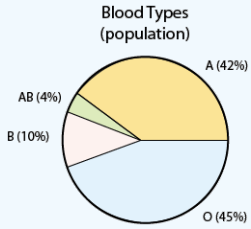
Learning Objectives

LO 6.20: Explain the concepts of sampling variability and sampling distribution.

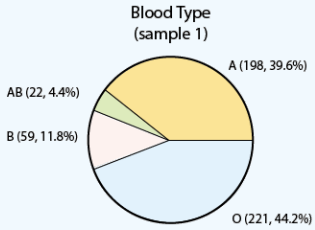
To better understand the relationship between sample and population, let's consider the two examples that were mentioned in the introduction.

EXAMPLE 1: Blood Type - Sampling Variability

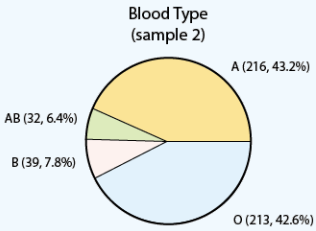
In the probability section, we presented the distribution of blood types in the entire U.S. **population**:



Assume now that we take a **sample** of 500 people in the United States, record their blood type, and display the sample results:



Note that the percentages (or proportions) that we found in our sample are slightly different than the population percentages. This is really not surprising. Since we took a sample of just 500, we cannot expect that our sample will behave exactly like the population, but if the sample is random (as it was), we expect to get results which are not that far from the population (as we did). If we took yet another sample of size 500:

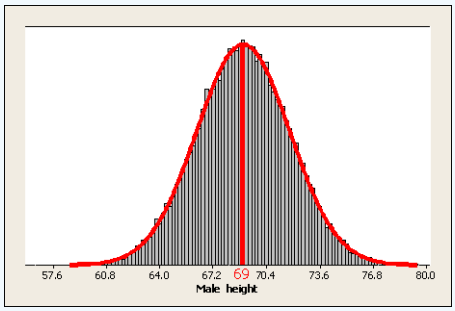


we again get sample results that are slightly different from the population figures, and also different from what we found in the first sample. This very intuitive idea, that sample results change from sample to sample, is called **sampling variability.**

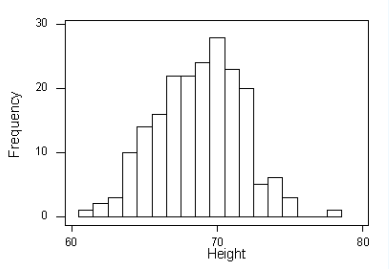
Let's look at another example:

EXAMPLE 2: Heights of Adults Males - Sampling Variability

Heights among the population of all adult males follow a normal distribution with a mean μ = mu =69 inches and a standard deviation σ = sigma =2.8 inches. Here is a probability display of this population distribution:

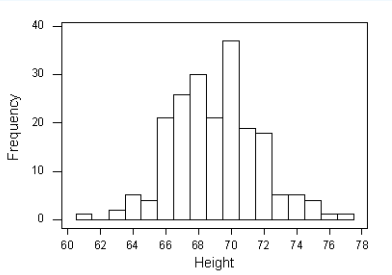


A sample of 200 males was chosen, and their heights were recorded. Here are the sample results:



The sample mean (x-bar) is 68.7 inches and the sample standard deviation (s) is 2.95 inches.

Again, note that the sample results are slightly different from the population. The histogram for this sample resembles the normal distribution, but is not as fine, and also the sample mean and standard deviation are slightly different from the population mean and standard deviation. Let's take another sample of 200 males:



The sample mean (x-bar) is 69.1 inches and the sample standard deviation (s) is 2.66 inches.

Again, as in Example 1 we see the idea of **sampling variability.** In this second sample, the results are pretty close to the population, but different from the results we found in the first sample.

In both the examples, we have numbers that describe the population, and numbers that describe the sample. In Example 1, the number 42% is the population proportion of blood type A, and 39.6% is the sample proportion (in sample 1) of blood type A. In Example 2, 69 and 2.8 are the population mean and standard deviation, and (in sample 1) 68.7 and 2.95 are the sample mean and standard deviation.

A **parameter** is a number that describes the population.

A **statistic** is a number that is computed from the sample.

✔ EXAMPLE 3: Parameters vs. Statistics from Example 1 and 2

In Example 1: 42% (0.42) is the parameter and 39.6% (0.396) is a statistic (and 43.2% is another statistic).

In Example 2: 69 and 2.8 are the parameters and 68.7 and 2.95 are statistics (69.1 and 2.66 are also statistics).

In this course, as in the examples above, we focus on the following parameters and statistics:

- population proportion and sample proportion
- population mean and sample mean
- population standard deviation and sample standard deviation

The following table summarizes the three pairs, and gives the notation

	(Population) Parameter	(Sample) Statistic
Proportion	p	$\hat{p}$
Mean	μ	$\bar{x}$
Standard Deviation	σ	s

The only new notation here is p for population proportion (p = 0.42 for type A in Example 1), and p-hat (using the "hat" symbol ∧ over the p) for the sample proportion which is 0.396 in Example 1, sample 1).

Comments:

- Parameters are usually unknown, because it is impractical or impossible to know exactly what values a variable takes for every member of the population.
- Statistics are computed from the sample, and vary from sample to sample due to **sampling variability**.

In the last part of the course, statistical inference, we will learn how to use a statistic to draw conclusions about an unknown parameter, either by estimating it or by deciding whether it is reasonable to conclude that the parameter equals a proposed value.

Now we'll learn about the behavior of the statistics assuming that we know the parameters. So, for example, if we know that the population proportion of blood type A in the population is 0.42, and we take a random sample of size 500, what do we expect the sample proportion p-hat to be? Specifically we ask:

- What is the distribution of all possible sample proportions from samples of size 500?
- Where is it centered?
- How much variation exists among different sample proportions from samples of size 500?
- How far off the true value of 0.42 might we expect to be?

Here are some more examples:

✔ EXAMPLE 4: Parameters vs. Statistics

If students picked numbers completely at random from the numbers 1 to 20, the proportion of times that the number 7 would be picked is 0.05. When 15 students picked a number "at random" from 1 to 20, 3 of them picked the number 7. Identify the parameter and accompanying statistic in this situation.

The parameter is the population proportion of random selections resulting in the number 7, which is p = 0.05. The accompanying statistic is the sample proportion (p-hat) of selections resulting in the number 7, which is 3/15=0.20.

Note: Unrelated to our current discussion, this is an interesting illustration of how we (humans) are not very good at doing things randomly. I used to ask a similar question in introductory statistics courses where I asked students to RANDOMLY pick a number between 1 and 10. The number of students choosing 7 is almost always MUCH larger than would be predicted if the results were truly random.

Try it with some of your friends and family and see if you get similar results. We really like the number 7! Interestingly, if students were aware of this phenomenon, then they tended to pick 3 most often. This is interesting since if choices were truly random, we should see a relatively equal proportion for each number :-)

✔ EXAMPLE 5: Parameters vs. Statistics

The length of human pregnancies has a mean of 266 days and a standard deviation of 16 days. A random sample of 9 pregnant women was observed to have a mean pregnancy length of 270 days, with a standard deviation of 14 days. Identify the parameters and accompanying statistics in this situation.

The parameters are population mean μ = mu =266 and population standard deviation σ = sigma = 16. The accompanying statistics are sample mean (x-bar) = 270 and sample standard deviation (s) = 14.

The first step to drawing conclusions about parameters based on the accompanying statistics is to understand how sample statistics behave relative to the parameter(s) that summarizes the entire population. We begin with the behavior of sample proportion relative

to population proportion (when the variable of interest is categorical). After that, we will explore the behavior of sample mean relative to population mean (when the variable of interest is quantitative).

> **Did I Get This?:** Parameters vs. Statistics

Sampling Distribution of the Sample Mean, x-bar
Sampling Distribution of the Sample Proportion, p-hat
Summary (Unit 3B - Sampling Distributions)

Sampling Distribution of the Sample Mean, x-bar

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Behavior of the Sample Mean (x-bar)

Learning Objectives

LO 6.22: Apply the sampling distribution of the sample mean as summarized by the Central Limit Theorem (when appropriate). In particular, be able to identify unusual samples from a given population.

So far, we've discussed the behavior of the statistic p-hat, the sample proportion, relative to the parameter p, the population proportion (when the variable of interest is categorical).

We are now moving on to explore the behavior of the statistic x-bar, the sample mean, relative to the parameter μ (mu), the population mean (when the variable of interest is quantitative).

Let's begin with an example.

EXAMPLE 9: Behavior of Sample Means

Birth weights are recorded for all babies in a town. The mean birth weight is 3,500 grams, μ = mu = 3,500 g. If we collect many random samples of 9 babies at a time, how do you think sample means will behave?

Here again, we are working with a random variable, since random samples will have means that vary unpredictably in the short run but exhibit patterns in the long run.

Based on our intuition and what we have learned about the behavior of sample proportions, we might expect the following about the distribution of sample means:

Center: Some sample means will be on the low side — say 3,000 grams or so — while others will be on the high side — say 4,000 grams or so. In repeated sampling, we might expect that the random samples will average out to the underlying population mean of 3,500 g. In other words, the mean of the sample means will be μ (mu), just as the mean of sample proportions was p.

Spread: For large samples, we might expect that sample means will not stray too far from the population mean of 3,500. Sample means lower than 3,000 or higher than 4,000 might be surprising. For smaller samples, we would be less surprised by sample means that varied quite a bit from 3,500. In others words, we might expect greater variability in sample means for smaller samples. So sample size will again play a role in the spread of the distribution of sample measures, as we observed for sample proportions.

Shape: Sample means closest to 3,500 will be the most common, with sample means far from 3,500 in either direction progressively less likely. In other words, the shape of the distribution of sample means should bulge in the middle and taper at the ends with a shape that is somewhat normal. This, again, is what we saw when we looked at the sample proportions.

Comment:

- The **distribution** of the values of the sample mean (x-bar) in repeated **samples** is called the **sampling distribution of x-bar**.

Let's look at a simulation:

Video

Video: Simulation #3 (x-bar) (4:31)

Did I Get This?: Simulation #3 (x-bar)

The results we found in our simulations are not surprising. Advanced probability theory confirms that by asserting the following:

The Sampling Distribution of the Sample Mean

If repeated random samples of a given size n are taken from a population of values for a quantitative variable, where the population mean is μ (mu) and the population standard deviation is σ (sigma) then the mean of all sample means (x-bars) is population mean μ (mu).

As for the spread of all sample means, theory dictates the behavior much more precisely than saying that there is less spread for larger samples. In fact, the standard deviation of all sample means is directly related to the sample size, n as indicated below.

$$\text{The standard deviation of all sample means } (\bar{x}) \text{ is exactly } \frac{\sigma}{\sqrt{n}}$$

Since the square root of sample size n appears in the denominator, the standard deviation does decrease as sample size increases.

Learn by Doing: Sampling Distribution (x-bar)

Let's compare and contrast what we now know about the sampling distributions for sample means and sample proportions.

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal IF np ≥ 10 and n(1 - p) ≥ 10
Quantitative (example: age)	μ = population mean, σ = population standard deviation	$\bar{x}$ = sample mean	μ	$\frac{\sigma}{\sqrt{n}}$	**When will the distribution of sample means be approximately normal ?**

Now we will investigate the shape of the sampling distribution of sample means. When we were discussing the sampling distribution of sample proportions, we said that this distribution is approximately normal if np ≥ 10 and n(1 – p) ≥ 10. In other words, we had a guideline based on sample size for determining the conditions under which we could use normal probability calculations for sample proportions.

When will the distribution of sample means be approximately normal? Does this depend on the size of the sample?

It seems reasonable that a population with a normal distribution will have sample means that are normally distributed even for very small samples. We saw this illustrated in the previous simulation with samples of size 10.

What happens if the distribution of the variable in the population is heavily skewed? Do sample means have a skewed distribution also? If we take really large samples, will the sample means become more normally distributed?

In the next simulation, we will investigate these questions.

Video

Video: Simulation #4 (x-bar) (5:02)

Did I Get This?: Simulation #4 (x-bar)

To summarize, the distribution of sample means will be approximately normal as long as the sample size is large enough. This discovery is probably the single most important result presented in introductory statistics courses. It is stated formally as the **Central Limit Theorem**.

We will depend on the Central Limit Theorem again and again in order to do normal probability calculations when we use sample means to draw conclusions about a population mean. We now know that we can do this even if the population distribution is not

normal.

How large a sample size do we need in order to assume that sample means will be normally distributed? Well, it really depends on the population distribution, as we saw in the simulation. The general rule of thumb is that samples of size 30 or greater will have a fairly normal distribution regardless of the shape of the distribution of the variable in the population.

Applet: Sampling Distribution for a Sample Mean

Comment:

- For categorical variables, our claim that sample proportions are approximately normal for large enough n is actually a special case of the Central Limit Theorem. In this case, we think of the data as 0's and 1's and the "average" of these 0's and 1's is equal to the proportion we have discussed.

Before we work some examples, let's compare and contrast what we now know about the sampling distributions for sample means and sample proportions.

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal if np ≥ 10 and n(1 - p) ≥ 10
Quantitative (example: age)	μ = population mean, σ = population standard deviation	$\bar{x}$ = sample mean	μ	$\frac{\sigma}{\sqrt{n}}$	Normal if n > 30 (always normal if population is normal)

Learn by Doing: Using the Sampling Distribution of x-bar

EXAMPLE 10: Using the Sampling Distribution of x-bar

Household size in the United States has a mean of 2.6 people and standard deviation of 1.4 people. It should be clear that this distribution is skewed right as the smallest possible value is a household of 1 person but the largest households can be very large indeed.

(a) What is the probability that a randomly chosen household has more than 3 people?

A normal approximation should not be used here, because the distribution of household sizes would be considerably skewed to the right. We do not have enough information to solve this problem.

(b) What is the probability that the mean size of a random sample of 10 households is more than 3?

By anyone's standards, 10 is a small sample size. The Central Limit Theorem does not guarantee sample mean coming from a skewed population to be approximately normal unless the sample size is large.

(c) What is the probability that the mean size of a random sample of 100 households is more than 3?

Now we may invoke the Central Limit Theorem: even though the distribution of household size X is skewed, the distribution of sample mean household size (x-bar) is approximately normal for a large sample size such as 100. Its mean is the same as the population mean, 2.6, and its standard deviation is the population standard deviation divided by the square root of the sample size:

$$\frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{100}} = 0.14$$

To find

$$P(\bar{x} > 3)$$

we standardize 3 to into a z-score by subtracting the mean and dividing the result by the standard deviation (of the sample mean). Then we can find the probability using the standard normal calculator or table.

$$P(\bar{x} > 3) = P\left(Z > \frac{3 - 2.6}{\frac{1.4}{\sqrt{100}}}\right) = P(Z > 2.86) = 0.0021$$

Households of more than 3 people are, of course, quite common, but it would be extremely unusual for the mean size of a sample of 100 households to be more than 3.

The purpose of the next activity is to give guided practice in finding the sampling distribution of the sample mean (x-bar), and use it to learn about the likelihood of getting certain values of x-bar.

Learn by Doing: Using the Sampling Distribution of x-bar #2

Did I Get This?: Using the Sampling Distribution of x-bar

Sampling Distribution of the Sample Proportion, p-hat

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Behavior of Sample Proportions

Learning Objectives

LO 6.21: Apply the sampling distribution of the sample proportion (when appropriate). In particular, be able to identify unusual samples from a given population.

EXAMPLE 6: Behavior of Sample Proportions

Approximately 60% of all part-time college students in the United States are female. (In other words, the population proportion of females among part-time college students is p = 0.6.) What would you expect to see in terms of the behavior of a sample proportion of females (p-hat) if random samples of size 100 were taken from the population of all part-time college students?

As we saw before, due to sampling variability, sample proportion in random samples of size 100 will take numerical values which vary according to the laws of chance: in other words, sample proportion is a **random variable**. To summarize the behavior of any random variable, we focus on three features of its distribution: the center, the spread, and the shape.

Based only on our intuition, we would expect the following:

Center: Some sample proportions will be on the low side — say, 0.55 or 0.58 — while others will be on the high side — say, 0.61 or 0.66. It is reasonable to expect all the sample proportions in repeated random samples to average out to the underlying population proportion, 0.6. In other words, the mean of the distribution of p-hat should be p.

Spread: For samples of 100, we would expect sample proportions of females not to stray too far from the population proportion 0.6. Sample proportions lower than 0.5 or higher than 0.7 would be rather surprising. On the other hand, if we were only taking samples of size 10, we would not be at all surprised by a sample proportion of females even as low as 4/10 = 0.4, or as high as 8/10 = 0.8. Thus, sample size plays a role in the spread of the distribution of sample proportion: there should be less spread for larger samples, more spread for smaller samples.

Shape: Sample proportions closest to 0.6 would be most common, and sample proportions far from 0.6 in either direction would be progressively less likely. In other words, the shape of the distribution of sample proportion should bulge in the middle and taper at the ends: it should be somewhat **normal.**

Comment:

- The **distribution** of the values of the sample proportions (p-hat) in repeated **samples** (of the same size) is called the **sampling distribution of p-hat**.

The purpose of the next video and activity is to check whether our intuition about the center, spread and shape of the sampling distribution of p-hat was correct via simulations.

Video

Video: Simulation #1 (p-hat) (4:13)

Did I Get This?: Simulation #1 (p-hat)

At this point, we have a good sense of what happens as we take random samples from a population. Our simulation suggests that our initial intuition about the shape and center of the sampling distribution is correct. If the population has a proportion of p, then random samples of the same size drawn from the population will have sample proportions close to p. More specifically, the distribution of sample proportions will have a mean of p.

We also observed that for this situation, the sample proportions are approximately normal. We will see later that this is not always the case. But if sample proportions are normally distributed, then the distribution is centered at p.

Now we want to use simulation to help us think more about the variability we expect to see in the sample proportions. Our intuition tells us that larger samples will better approximate the population, so we might expect less variability in large samples.

In the next walk-through we will use simulations to investigate this idea. After that walk-through, we will tie these ideas to more formal theory.

Video

Video: Simulation #2 (p-hat) (4:55)

Did I Get This?: Simulation #2 (p-hat)

The simulations reinforced what makes sense to our intuition. Larger random samples will better approximate the population proportion. When the sample size is large, sample proportions will be closer to p. In other words, the sampling distribution for large samples has less variability. Advanced probability theory confirms our observations and gives a more precise way to describe the standard deviation of the sample proportions. This is described next.

The Sampling Distribution of the Sample Proportion

If repeated random samples of a given size n are taken from a population of values for a categorical variable, where the proportion in the category of interest is p, then the mean of all sample proportions (p-hat) is the population proportion (p).

As for the spread of all sample proportions, theory dictates the behavior much more precisely than saying that there is less spread for larger samples. In fact, the standard deviation of all sample proportions is directly related to the sample size, n as indicated below.

$$\text{The standard deviation of all sample proportions } (\hat{p}) \text{ is exactly } \sqrt{\frac{p(1-p)}{n}}$$

Since the sample size n appears in the denominator of the square root, the standard deviation does decrease as sample size increases. Finally, the shape of the distribution of p-hat will be approximately normal as long as the sample size n is large enough. The convention is to require both np and n(1 – p) to be at least 10.

We can summarize all of the above by the following:

$$\hat{p} \text{ is normally distributed with a mean of } \mu_{\hat{p}} = p$$
$$\text{and a standard deviation } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$
$$\text{as long as } np \geq 10 \text{ and } n(1-p) \geq 10$$

Let’s apply this result to our example and see how it compares with our simulation.

In our example, n = 25 (sample size) and p = 0.6. Note that np = 15 ≥ 10 and n(1 – p) = 10 ≥ 10. Therefore we can conclude that p-hat is approximately a normal distribution with mean p = 0.6 and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{25}} = 0.097$$

(which is very close to what we saw in our simulation).

Comment:

- These results are similar to those for binomial random variables (X) discussed previously. Be careful not to confuse the results for the mean and standard deviation of X with those of p-hat.

Learn by Doing: Sampling Distribution of p-hat

Did I Get This?: Sampling Distribution of p-hat

If a sampling distribution is normally shaped, then we can apply the Standard Deviation Rule and use z-scores to determine probabilities. Let's look at some examples.

EXAMPLE 7: Using the Sample Distribution of p-hat

A random sample of 100 students is taken from the population of all part-time students in the United States, for which the overall proportion of females is 0.6.

(a) There is a 95% chance that the sample proportion (p-hat) falls between what two values?

First note that the distribution of p-hat has mean p = 0.6, standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.05$$

and a shape that is close to normal, since np = 100(0.6) = 60 and n(1 – p) = 100(0.4) = 40 are both greater than 10. The Standard Deviation Rule applies: the probability is approximately 0.95 that p-hat falls within 2 standard deviations of the mean, that is, between 0.6 – 2(0.05) and 0.6 + 2(0.05). There is roughly a 95% chance that p-hat falls in the interval (0.5, 0.7) for samples of this size.

(b) What is the probability that sample proportion p-hat is less than or equal to 0.56?

To find

$$P(\hat{p} \leq 0.56)$$

we standardize 0.56 into a z-score by subtracting the mean and dividing the result by the standard deviation. Then we can find the probability using the standard normal calculator or table.

$$P(\hat{p} \leq 0.56) = P\left(Z \leq \frac{0.56-0.6}{0.05}\right) = P(Z \leq -0.80) = 0.2119$$

To see the impact of the sample size on these probability calculations, consider the following variation of our example.

EXAMPLE 8: Using the Sample Distribution of p-hat

A random sample of **2500** students is taken from the population of all part-time students in the United States, for which the overall proportion of females is 0.6.

(a) There is a 95% chance that the sample proportion (p-hat) falls between what two values?

First note that the distribution of p-hat has mean p = 0.6, standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{2500}} = 0.01$$

and a shape that is close to normal, since np = 2500(0.6) = 1500 and n(1 – p) = 2500(0.4) = 1000 are both greater than 10. The Standard Deviation Rule applies: the probability is approximately 0.95 that p-hat falls within 2 standard deviations of the mean, that is, between 0.6 – 2(0.01) and 0.6 + 2(0.01). There is roughly a 95% chance that p-hat falls in the interval (0.58, 0.62) for samples of this size.

(b) What is the probability that sample proportion p-hat is less than or equal to 0.56?

To find

$$P(\hat{p} \leq 0.56)$$

we standardize 0.56 to into a z-score by subtracting the mean and dividing the result by the standard deviation. Then we can find the probability using the standard normal calculator or table.

$$P(\hat{p} \leq 0.56) = P\left(Z \leq \frac{0.56-0.6}{0.01}\right) = P(Z \leq -4) \approx 0$$

Comment:

- As long as the sample is truly random, the distribution of p-hat is centered at p, no matter what size sample has been taken. Larger samples have less spread. Specifically, when we multiplied the sample size by 25, increasing it from 100 to 2,500, the standard deviation was reduced to 1/5 of the original standard deviation. Sample proportion strays less from population proportion 0.6 when the sample is larger: it tends to fall anywhere between 0.5 and 0.7 for samples of size 100, whereas it tends to fall between 0.58 and 0.62 for samples of size 2,500. It is not so improbable to take a value as low as 0.56 for samples of 100 (probability is more than 20%) but it is almost impossible to take a value as low as 0.56 for samples of 2,500 (probability is virtually zero).

> **Applet:** Sampling Distribution for a Sample Proportion

Summary (Unit 3B - Sampling Distributions)

We have finally reached the end our discussion of probability with our discussion of sampling distributions, which can be viewed in two ways. On the one hand Sampling Distributions can be viewed as a special case of Random Variables since we discussed two special random variables: the sample mean (x-bar) and the sample proportion (p-hat). On the other hand, Sampling Distributions can be viewed as the bridge that takes us from probability to statistical inference.

As mentioned in the introduction, this last concept in probability is the bridge between the probability section and inference. It focuses on the relationship between sample values (**statistics**) and population values (**parameters**). Statistics vary from sample to sample due to **sampling variability**, and therefore can be regarded as **random variables** whose distribution we call the **sampling distribution**.

In our discussion of sampling distributions, we focused on two statistics, the **sample proportion**, p-hat and the **sample mean**, x-bar. Our goal was to explore the sampling distribution of these two statistics relative to their respective population parameters, p and μ (mu), and we found in **both** cases that under certain conditions the **sampling distribution is approximately normal**. This result is known as the **Central Limit Theorem.** As we'll see in the next section, the Central Limit Theorem is the foundation for statistical inference.

> **Outside Reading:** Little Handbook – Behavior of Sample Means (≈ 3000 words)

Sampling Distributions

A **parameter** is a number that describes the population, and a **statistic** is a number that describes the sample.

- Parameters are fixed, and in practice, usually unknown.
- Statistics change from sample to sample due to sampling variability.
- The behavior of the possible values the statistic can take in repeated samples is called the **sampling distribution** of that statistic.
- The following table summarizes the important information about the two sampling distributions we covered. Both of these results follow from the **central limit theorem** which basically states that as the sample size increases, the distribution of the average from a sample of size n becomes increasingly normally distributed.

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal if np ≥ 10 and n(1 - p) ≥ 10
Quantitative (example: age)	μ = population mean, σ = population standard deviation	$\bar{x}$ = sample mean	μ	$\frac{\sigma}{\sqrt{n}}$	Normal if n > 30 (always normal if population is normal)

CHAPTER OVERVIEW

Unit 4A: Introduction to Statistical Inference

CO-1: Describe the roles biostatistics serves in the discipline of public health.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Review: We are about to move into the inference component of the course and it is a good time to be sure you understand the basic ideas presented regarding exploratory data analysis.

- What is Data?
- Types of Variables
- Unit 1: Exploratory Data Analysis
  - One Categorical Variable
  - One Quantitative Variable
    - Histograms & Stemplots
    - Describing Distributions
    - Measures of Center
    - Measures of Spread
    - Measures of Position
    - Outliers
    - Boxplots
    - The “Normal” Shape

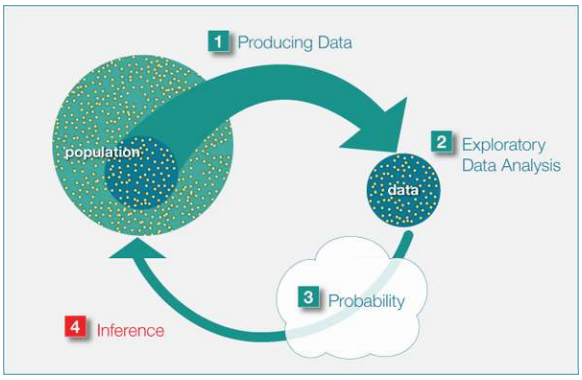
Video

Video: Unit 4A: Introduction to Statistical Inference (15:45)

Recall again the Big Picture, the four-step process that encompasses statistics: data production, exploratory data analysis, probability and inference.

We are about to start the fourth and final unit of this course, where we draw on principles learned in the other units (Exploratory Data Analysis, Producing Data, and Probability) in order to accomplish what has been our ultimate goal all along: use a sample to infer (or draw conclusions) about the population from which it was drawn.

As you will see in the introduction, the specific form of inference called for depends on the type of variables involved — either a single categorical or quantitative variable, or a combination of two variables whose relationship is of interest.



Introduction

Learning Objectives

LO 6.23: Explain how the concepts covered in Units 1 – 3 provide the basis for statistical inference.

We are about to start the fourth and final part of this course — statistical inference, where we draw conclusions about a population based on the data obtained from a sample chosen from it.

The purpose of this introduction is to review how we got here and how the previous units fit together to allow us to make reliable inferences. Also, we will introduce the various forms of statistical inference that will be discussed in this unit, and give a general outline of how this unit is organized.

In the **Exploratory Data Analysis** unit, we learned to display and summarize data that were obtained from a sample. Regardless of whether we had one variable and we examined its distribution, or whether we had two variables and we examined the relationship between them, it was always understood that these summaries applied **only** to the data at hand; we did not attempt to make claims about the larger population from which the data were obtained.

Such generalizations were, however, a long-term goal from the very beginning of the course. For this reason, in the unit on **Producing Data**, we took care to establish principles of sampling and study design that would be essential in order for us to claim that, to some extent, what is true for the sample should be also true for the larger population from which the sample originated.

These principles should be kept in mind throughout this unit on statistical inference, since the results that we will obtain will not hold if there was bias in the sampling process, or flaws in the study design under which variables' values were measured.

Perhaps the most important principle stressed in the Producing Data unit was that of randomization. Randomization is essential, not only because it prevents bias, but also because it permits us to rely on the laws of probability, which is the scientific study of random behavior.

In the **Probability** unit, we established basic laws for the behavior of random variables. We ultimately focused on two random variables of particular relevance: the sample mean (x-bar) and the sample proportion (p-hat), and the last section of the Probability unit was devoted to exploring their sampling distributions.

We learned what probability theory tells us to expect from the values of the sample mean and the sample proportion, given that the corresponding population parameters — the population mean (mu, μ) and the population proportion (p) — are known.

As we mentioned in that section, the value of such results is more theoretical than practical, since in real-life situations we seldom know what is true for the entire population. All we know is what we see in the sample, and we want to use this information to say something concrete about the larger population.

Probability theory has set the stage to accomplish this: learning what to expect from the value of the sample mean, given that population mean takes a certain value, teaches us (as we'll soon learn) what to expect from the value of the unknown population mean, given that a particular value of the sample mean has been observed.

Similarly, since we have established how the sample proportion behaves relative to population proportion, we will now be able to turn this around and say something about the value of the population proportion, based on an observed sample proportion. This process — inferring something about the population based on what is measured in the sample — is (as you know) called **statistical inference**.

Types of Inference

Learning Objectives

LO: 1.9 Distinguish between situations using a point estimate, an interval estimate, or a hypothesis test.

We will introduce three forms of statistical inference in this unit, each one representing a different way of using the information obtained in the sample to draw conclusions about the population. These forms are:

- Point Estimation
- Interval Estimation

- Hypothesis Testing

Obviously, each one of these forms of inference will be discussed at length in this section, but it would be useful to get at least an intuitive sense of the nature of each of these inference forms, and the difference between them in terms of the types of conclusions they draw about the population based on the sample results.

Point Estimation

In **point estimation**, we estimate an unknown parameter using a **single number** that is calculated from the sample data.

✔ EXAMPLE:

Based on sample results, we estimate that p, the proportion of all U.S. adults who are in favor of stricter gun control, is 0.6.

Interval Estimation

In **interval estimation**, we estimate an unknown parameter using an **interval of values** that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter).

✔ EXAMPLE:

Based on sample results, we are 95% confident that p, the proportion of all U.S. adults who are in favor of stricter gun control, is between 0.57 and 0.63.

Hypothesis Testing

In **hypothesis testing**, we begin with a claim about the population (we will call the null hypothesis), and we check **whether or not the data** obtained from the sample **provide evidence AGAINST this claim.**

✔ EXAMPLE:

It was claimed that among all U.S. adults, about half are in favor of stricter gun control and about half are against it. In a recent poll of a random sample of 1,200 U.S. adults, 60% were in favor of stricter gun control. This data, therefore, provides some evidence against the claim.

Soon we will determine the **probability** that we could have seen such a result (60% in favor) or more extreme **IF** in fact the true proportion of all U.S. adults who favor stricter gun control is actually 0.5 (the value in the claim the data attempts to refute).

✔ EXAMPLE:

It is claimed that among drivers 18-23 years of age (our population) there is no relationship between drunk driving and gender.

A roadside survey collected data from a random sample of 5,000 drivers and recorded their gender and whether they were drunk.

The collected data showed roughly the same percent of drunk drivers among males and among females. These data, therefore, do not give us any reason to reject the claim that there is no relationship between drunk driving and gender.

Did I Get This?: Types of Inference

In terms of organization, the Inference unit consists of two main parts: Inference for One Variable and Inference for Relationships between Two Variables. The organization of each of these parts will be discussed further as we proceed through the unit.

Inference for One Variable

The next two topics in the inference unit will deal with inference for one variable. Recall that in the Exploratory Data Analysis (EDA) unit, when we learned about summarizing the data obtained from one variable where we learned about examining distributions, we distinguished between two cases; categorical data and quantitative data.

We will make a similar distinction here in the inference unit. In the EDA unit, the type of variable determined the displays and numerical measures we used to summarize the data. In Inference, the type of variable of interest (categorical or quantitative) will determine what population parameter is of interest.

- When the variable of interest is **categorical**, the population parameter that we will infer about is the **population proportion (p)** associated with that variable. For example, if we are interested in studying opinions about the death penalty among U.S. adults, and thus our variable of interest is "death penalty (in favor/against)," we'll choose a sample of U.S. adults and use the collected data to make an inference about p, the proportion of U.S. adults who support the death penalty.
- When the variable of interest is **quantitative**, the population parameter that we infer about is the **population mean (mu, μ)** associated with that variable. For example, if we are interested in studying the annual salaries in the population of teachers in a certain state, we'll choose a sample from that population and use the collected salary data to make an inference about μ, the mean annual salary of all teachers in that state.

The following outlines describe some of the important points about the process of inferential statistics as well as compare and contrast how researchers and statisticians approach this process.

Outline of Process of Inference

Here is another restatement of the big picture of statistical inference as it pertains to the two simple examples we will discuss first.

- A simple random sample is taken from a population of interest.
- In order to estimate a **population parameter**, a **statistic** is calculated from the **sample**. For example:

Sample mean (x-bar)

Sample proportion (p-hat)

- We then learn about the **DISTRIBUTION** of this statistic in **repeated sampling (theoretically)**. We now know these are called **sampling distributions**!
- Using THIS sampling distribution we can make **inferences** about our **population parameter** based upon our **sample statistic**.

It is this last step of statistical inference that we are interested in discussing now.

Applied Steps (What do researchers do?)

One issue for students is that the theoretical process of statistical inference is only a small part of the applied steps in a research project. Previously, in our discussion of the role of biostatistics, we defined these steps to be:

1. Planning/design of study
2. Data collection
3. Data analysis
4. Presentation
5. Interpretation

You can see that:

- **Both exploratory data analysis** and **inferential methods** will fall into the category of **"Data Analysis"** in our previous list.

- **Probability is hiding** in the applied steps in the form of **probability sampling plans, estimation of desired probabilities,** and **sampling distributions.**

Among researchers, the following represent some of the important questions to address when conducting a study.

- What is the population of interest?
- What is the question or statistical problem?
- How to sample to best address the question given the available resources?
- How to analyze the data?
- How to report the results?

AFTER you know what you are going to do, then you can begin collecting data!

Theoretical Steps (What do statisticians do?)

Statisticians, on the other hand, need to ask questions like these:

- What **assumptions** can be reasonably made about the **population**?
- What **parameter(s)** in the **population** do we need to **estimate** in order to address the research question?
- What **statistic(s)** from our **sample** data can be used to **estimate** the **unknown parameter(s)**?
- How does each **statistic behave**?
  - Is it **unbiased**?
  - How **variable** will it be for the planned sample size?
  - What is the **distribution** of this statistic? (Sampling Distribution)

Then, we will see that we can use the sampling distribution of a statistic to:

- Provide **confidence interval estimates** for the corresponding **parameter**.
- Conduct **hypothesis tests** about the corresponding **parameter**.

Standard Error of a Statistic

Learning Objectives

LO: 1.10: Define the standard error of a statistic precisely and relate it to the concept of the sampling distribution of a statistic.

In our discussion of sampling distributions, we discussed the **variability of sample statistics**; here is a quick review of this general concept and a formal **definition of the standard error of a statistic**.

- All statistics calculated from samples are **random variables.**
- The distribution of a statistic (from a sample of a given sample size) is called the **sampling distribution of the statistic.**
- The **standard deviation of the sampling distribution** of a particular statistic is called the **standard error of the statistic** and measures variability of the statistic for a particular sample size.

The **standard error** of a statistic is the **standard deviation of the sampling distribution of that statistic**, where the sampling distribution is defined as the distribution of a particular statistic in repeated sampling.

- The standard error is an extremely common measure of the variability of a sample statistic.

EXAMPLE:

In our discussion of sampling distributions, we looked at a situation involving a random sample of 100 students taken from the population of all part-time students in the United States, for which the overall proportion of females is 0.6. Here we have a categorical variable of interest, gender.

We determined that the distribution of all possible values of p-hat (that we could obtain for repeated simple random samples of this size from this population) has mean p = 0.6 and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.05$$

which we have now learned is more formally called the standard error of p-hat. **In this case, the true standard error of p-hat will be 0.05**.

We also showed how we can use this information along with information about the center (mean or expected value) to calculate probabilities associated with particular values of p-hat. For example, what is the probability that sample proportion p-hat is less than or equal to 0.56? After verifying the sample size requirements are reasonable, we can use a normal distribution to approximate

$$P(\hat{p} \leq 0.56) = P\left(Z \leq \frac{0.56-0.6}{0.05}\right) = P(Z \leq -0.80) = 0.2119$$

EXAMPLE:

Similarly, for a quantitative variable, we looked at an example of household size in the United States which has a mean of 2.6 people and standard deviation of 1.4 people.

If we consider taking a simple random sample of 100 households, we found that the distribution of sample means (x-bar) is approximately normal for a large sample size such as n = 100.

The sampling distribution of x-bar has a mean which is the same as the population mean, 2.6, and its standard deviation is the population standard deviation divided by the square root of the sample size:

$$\frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{100}} = 0.14$$

Again, this standard deviation of the sampling distribution of x-bar is more commonly called the **standard error of x-bar**, in this case 0.14. And we can use this information (the center and spread of the sampling distribution) to find probabilities involving particular values of x-bar.

$$P(\bar{x} > 3) = P\left(Z > \frac{3-2.6}{\frac{1.4}{\sqrt{100}}}\right) = P(Z > 2.86) = 0.0021$$

Estimation
Hypothesis Testing
Wrap-Up (Inference for One Variable)

Estimation

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Video: Estimation (11:40)

Introduction

In our Introduction to Inference we defined point estimates and interval estimates.

- In **point estimation**, we estimate an unknown parameter using a single number that is calculated from the sample data.
- In **interval estimation**, we estimate an unknown parameter using an interval of values that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter).

In this section, we will introduce the concept of a confidence interval and learn to calculate confidence intervals for population means and population proportions (when certain conditions are met).

In Unit 4B, we will see that confidence intervals are useful whenever we wish to use data to estimate an unknown population parameter, even when this parameter is estimated using multiple variables (such as our cases: CC, CQ, QQ).

For example, we can construct confidence intervals for the slope of a regression equation or the correlation coefficient. In doing so we are always using our data to provide an interval estimate for an unknown population parameter (the TRUE slope, or the TRUE correlation coefficient).

Point Estimation

Learning Objectives

LO 4.29: Determine and use the correct point estimates for specified population parameters.

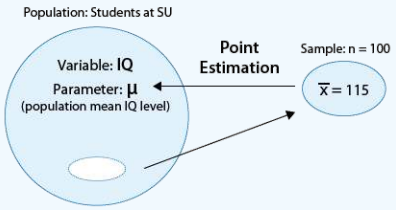
Point estimation is the form of statistical inference in which, based on the sample data, we estimate the unknown parameter of interest using a **single** value (hence the name **point** estimation). As the following two examples illustrate, this form of inference is quite intuitive.

EXAMPLE:

Suppose that we are interested in studying the IQ levels of students at Smart University (SU). In particular (since IQ level is a quantitative variable), we are interested in estimating μ (mu), the mean IQ level of all the students at SU.

A random sample of 100 SU students was chosen, and their (sample) mean IQ level was found to be 115 (x-bar).

If we wanted to estimate μ (mu), the population mean IQ level, by a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample mean which is 115. We say that 115 is the **point estimate** for μ (mu), and in general, we'll always use the sample mean (x-bar) as the **point estimator** for μ (mu). (Note that when we talk about the **specific** value (115), we use the term **estimate**, and when we talk in general about the **statistic** x-bar, we use the term **estimator**. The following figure summarizes this example:

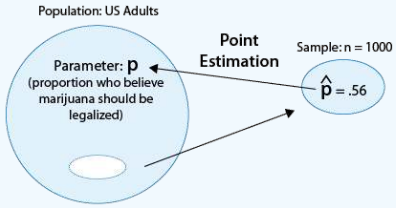


Here is another example.

EXAMPLE:

Suppose that we are interested in the opinions of U.S. adults regarding legalizing the use of marijuana. In particular, we are interested in the parameter p, the proportion of U.S. adults who believe marijuana should be legalized.

Suppose a poll of 1,000 U.S. adults finds that 560 of them believe marijuana should be legalized. If we wanted to estimate p, the population proportion, using a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample proportion p-hat = 560/1000 = 0.56. We say in this case that 0.56 is the **point estimate** for p, and in general, we'll always use p-hat as the **point estimator** for p. (Note, again, that when we talk about the **specific value** (0.56), we use the term **estimate**, and when we talk in general about the **statistic** p-hat, we use the term **estimator**. Here is a visual summary of this example:

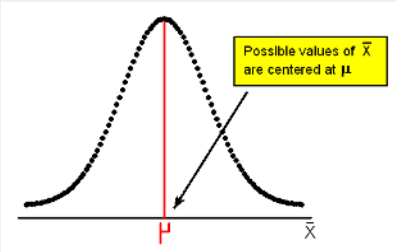


Did I Get This?: Point Estimation

Desired Properties of Point Estimators

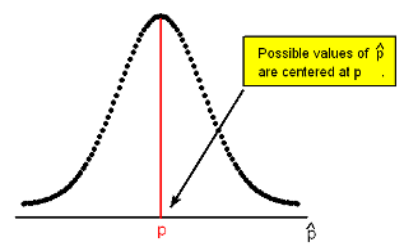
You may feel that since it is so intuitive, you could have figured out point estimation on your own, even without the benefit of an entire course in statistics. Certainly, our intuition tells us that the best estimator for the population mean (mu, μ) should be x-bar, and the best estimator for the population proportion p should be p-hat.

Probability theory does more than this; it actually gives an explanation (beyond intuition) **why** x-bar and p-hat are the good choices as point estimators for μ (mu) and p, respectively. In the Sampling Distributions section of the Probability unit, we learned about the sampling distribution of x-bar and found that **as long as a sample is taken at random**, the distribution of sample means is exactly centered at the value of population mean.



Our statistic, x-bar, is therefore said to be an **unbiased** estimator for μ (mu). Any particular sample mean might turn out to be less than the actual population mean, or it might turn out to be more. But in the long run, such sample means are "on target" in that they will not underestimate any more or less often than they overestimate.

Likewise, we learned that the sampling distribution of the sample proportion, p-hat, is centered at the population proportion p (as long as the sample is taken at random), thus making p-hat an unbiased estimator for p.



As stated in the introduction, probability theory plays an essential role as we establish results for statistical inference. Our assertion above that sample mean and sample proportion are unbiased estimators is the first such instance.

Importance of Sampling and Design

Notice how important the principles of sampling and design are for our above results: if the sample of U.S. adults in (example 2 on the previous page) was not random, but instead included predominantly college students, then 0.56 would be a biased estimate for p, the proportion of all U.S. adults who believe marijuana should be legalized.

If the survey design were flawed, such as loading the question with a reminder about the dangers of marijuana leading to hard drugs, or a reminder about the benefits of marijuana for cancer patients, then 0.56 would be biased on the low or high side, respectively.

Caution

Our point estimates are truly **unbiased** estimates for the population parameter **only if the sample is random and the study design is not flawed**.

Standard Error and Sample Size

Not only are the sample mean and sample proportion on target as long as the samples are random, but **their precision improves as sample size increases**.

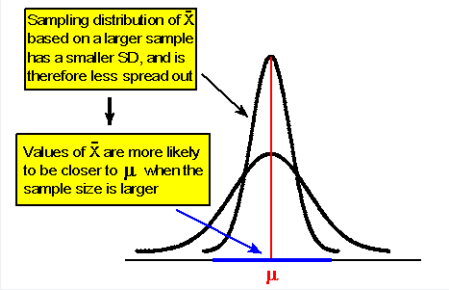
Again, there are two “layers” here for explaining this.

Intuitively, larger sample sizes give us more information with which to pin down the true nature of the population. We can therefore expect the sample mean and sample proportion obtained from a larger sample to be closer to the population mean and proportion, respectively. In the extreme, when we sample the whole population (which is called a census), the sample mean and sample proportion will exactly coincide with the population mean and population proportion.There is another layer here that, again, comes from what we learned about the sampling distributions of the sample mean and the sample proportion. Let’s use the sample mean for the explanation.

Recall that the sampling distribution of the sample mean x-bar is, as we mentioned before, centered at the population mean μ (mu)and has a standard error (standard deviation of the statistic, x-bar) of

$$\text{standard deviation of } \frac{\sigma}{\sqrt{n}}$$

As a result, as the sample size n increases, the sampling distribution of x-bar gets less spread out. This means that values of x-bar that are based on a larger sample are more likely to be closer to μ (mu) (as the figure below illustrates):



Similarly, since the sampling distribution of p-hat is centered at p and has a

$$\text{standard deviation of } \sqrt{\frac{p(1-p)}{n}}$$

which decreases as the sample size gets larger, values of p-hat are more likely to be closer to p when the sample size is larger.

Another Point Estimator

Another example of a point estimator is using sample standard deviation,

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

to estimate population standard deviation, σ (sigma).

In this course, we will not be concerned with estimating the population standard deviation for its own sake, but since we will often substitute the sample standard deviation (s) for σ (sigma) when standardizing the sample mean, it is worth pointing out that **s is an unbiased estimator for σ** (sigma).

If we had divided by n instead of n – 1 in our estimator for population standard deviation, then in the long run our sample variance would be guilty of a slight underestimation. Division by n – 1 accomplishes the goal of making this point estimator unbiased.

The reason that our formula for s, introduced in the Exploratory Data Analysis unit, involves division by n – 1 instead of by n is the fact that we wish to use unbiased estimators in practice.

Let's Summarize

- We use p-hat (sample proportion) as a point estimator for p (population proportion). It is an unbiased estimator: its long-run distribution is centered at p as long as the sample is random.
- We use x-bar (sample mean) as a point estimator for μ (mu, population mean). It is an unbiased estimator: its long-run distribution is centered at μ (mu) as long as the sample is random.
- In both cases, the larger the sample size, the more precise the point estimator is. In other words, the larger the sample size, the more likely it is that the sample mean (proportion) is close to the unknown population mean (proportion).

Did I Get This?: Properties of Point Estimators

Interval Estimation

Point estimation is simple and intuitive, but also a bit problematic. Here is why:

When we estimate μ (mu) by the sample mean x-bar we are almost guaranteed to make some kind of error. Even though we know that the values of x-bar fall around μ (mu), it is very unlikely that the value of x-bar will fall exactly at μ (mu).

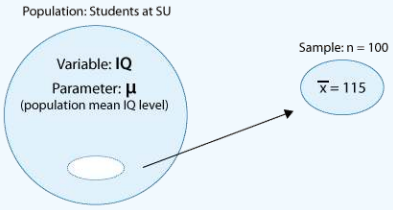
Given that such errors are a fact of life for point estimates (by the mere fact that we are basing our estimate on one sample that is a small fraction of the population), these estimates are in themselves of limited usefulness, unless we are able to quantify the extent of the estimation error. Interval estimation addresses this issue. The idea behind **interval estimation** is, therefore, to enhance the simple point estimates by supplying information about the size of the error attached.

In this introduction, we'll provide examples that will give you a solid intuition about the basic idea behind interval estimation.

EXAMPLE:

Consider the example that we discussed in the point estimation section:

Suppose that we are interested in studying the IQ levels of students attending Smart University (SU). In particular (since IQ level is a quantitative variable), we are interested in estimating μ (mu), the mean IQ level of all the students in SU. A random sample of 100 SU students was chosen, and their (sample) mean IQ level was found to be 115 (x-bar).



In point estimation we used x-bar = 115 as the point estimate for μ (mu). However, we had no idea of what the estimation error involved in such an estimation might be. Interval estimation takes point estimation a step further and says something like:

"I am 95% confident that by using the point estimate x-bar = 115 to estimate μ (mu), I am off by no more than 3 IQ points. In other words, I am 95% confident that μ (mu) is within 3 of 115, or between 112 (115 – 3) and 118 (115 + 3)."

Yet another way to say the same thing is: I am 95% confident that μ (mu) is somewhere in (or covered by) the interval (112,118). (**Comment:** At this point you should not worry about, or try to figure out, how we got these numbers. We'll do that later. All we want to do here is make sure you understand the idea.)

Note that while point estimation provided just one number as an estimate for μ (mu) of 115, interval estimation provides a whole interval of "plausible values" for μ (mu) (between 112 and 118), and also attaches the level of our confidence that this interval indeed includes the value of μ (mu) to our estimation (in our example, 95% confidence). The interval (112,118) is therefore called "a 95% confidence interval for μ (mu)."

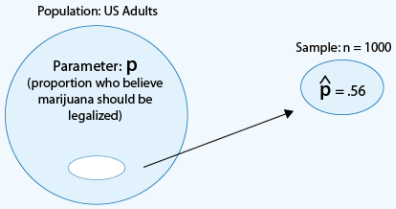
Let's look at another example:

EXAMPLE:

Let's consider the second example from the point estimation section.

Suppose that we are interested in the opinions of U.S. adults regarding legalizing the use of marijuana. In particular, we are interested in the parameter p, the proportion of U.S. adults who believe marijuana should be legalized.

Suppose a poll of 1,000 U.S. adults finds that 560 of them believe marijuana should be legalized.



If we wanted to estimate p, the population proportion, by a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample proportion p-hat = 560/1000=0.56.

Interval estimation would take this a step further and say something like:

"I am 90% confident that by using 0.56 to estimate the true population proportion, p, I am off by (or, I have an error of) no more than 0.03 (or 3 percentage points). In other words, I am 90% confident that the actual value of p is somewhere between 0.53 (0.56 – 0.03) and 0.59 (0.56 + 0.03)."

Yet another way of saying this is: "I am 90% confident that p is covered by the interval (0.53, 0.59)."

In this example, (0.53, 0.59) is a 90% confidence interval for p.

Let's summarize

The two examples showed us that the idea behind interval estimation is, instead of providing just one number for estimating an unknown parameter of interest, to provide an interval of plausible values of the parameter plus a level of confidence that the value of the parameter is covered by this interval.

We are now going to go into more detail and learn how these confidence intervals are created and interpreted in context. As you'll see, the ideas that were developed in the "Sampling Distributions" section of the Probability unit will, again, be very important. Recall that for point estimation, our understanding of sampling distributions leads to verification that our statistics are unbiased and gives us a precise formulas for the standard error of our statistics.

We'll start by discussing confidence intervals for the population mean μ (mu), and later discuss confidence intervals for the population proportion p.

Population Means (Part 1)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.30: Interpret confidence intervals for population parameters in context.

Learning Objectives

LO 4.31: Find confidence intervals for the population mean using the normal distribution (Z) based confidence interval formula (when required conditions are met) and perform sample size calculations.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.24: Explain the connection between the sampling distribution of a statistic, and its properties as a point estimator.

Learning Objectives

LO 6.25: Explain what a confidence interval represents and determine how changes in sample size and confidence level affect the precision of the confidence interval.

Video

Video: Population Means – Part 1 (11:14)

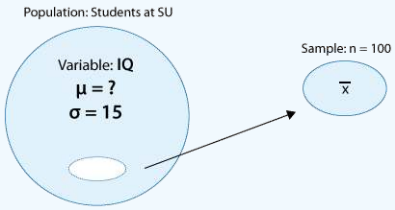
As the introduction mentioned, we'll start our discussion on interval estimation with interval estimation for the population mean μ (mu). We'll start by showing how a 95% confidence interval is constructed, and later generalize to other levels of confidence. We'll also discuss practical issues related to interval estimation.

Recall the IQ example:

EXAMPLE:

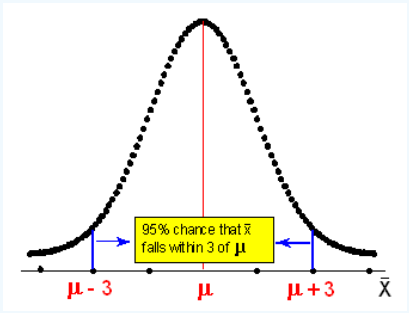
Suppose that we are interested in studying the IQ levels of students at Smart University (SU). In particular (since IQ level is a quantitative variable), we are interested in estimating μ (mu), the mean IQ level of all the students at SU.

We will assume that from past research on IQ scores in different universities, it is known that the IQ standard deviation in such populations is σ (sigma) = 15. In order to estimate μ (mu), a random sample of 100 SU students was chosen, and their (sample) mean IQ level is calculated (let's assume, for now, that we have not yet found the sample mean).



We will now show the rationale behind constructing a 95% confidence interval for the population mean μ (mu).

- We learned in the "Sampling Distributions" section of probability that according to the central limit theorem, the sampling distribution of the sample mean x-bar is approximately normal with a mean of μ (mu) and standard deviation of σ/sqrt(n) = sigma/sqrt(n). In our example, then, (where σ (sigma) = 15 and n = 100), the possible values of x-bar, the sample mean IQ level of 100 randomly chosen students, is approximately normal, with mean μ (mu) and standard deviation 15/sqrt(100) = 1.5.
- Next, we recall and apply the Standard Deviation Rule for the normal distribution, and in particular its second part: There is a 95% chance that the sample mean we will find in our sample falls within 2 * 1.5 = 3 of μ (mu).



Obviously, if there is a certain distance between the sample mean and the population mean, we can describe that distance by starting at either value. So, if the sample mean (x-bar) falls within a certain distance of the population mean μ (mu), then the population mean μ (mu) falls within the same distance of the sample mean.

Therefore, the statement, "There is a 95% **chance** that the **sample** mean x-bar falls within 3 units of μ (mu)" can be rephrased as: "We are 95% **confident** that the **population** mean μ (mu) falls within 3 units of the x-bar we found in our sample."

So, if we happen to get a sample mean of x-bar = 115, then we are 95% confident that μ (mu) falls within 3 units of 115, or in other words that μ (mu) is covered by the interval (115 – 3, 115 + 3) = (112,118).

(On later pages, we will use similar reasoning to develop a general formula for a confidence interval.)

Comment:

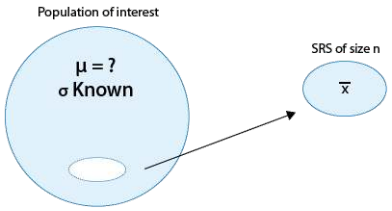
- Note that the first phrasing is about x-bar, which is a random variable; that's why it makes sense to use probability language. But the second phrasing is about μ (mu), which is a parameter, and thus is a "fixed" value that does not change, and that's why we should not use probability language to discuss it. In these problems, it is our x-bar that will change when we repeat the process, not μ (mu). This point will become clearer after you do the activities which follow.

The General Case

Let's generalize the IQ example. Suppose that we are interested in estimating the unknown population mean (μ, mu) based on a random sample of size n. Further, we assume that the population standard deviation (σ, sigma) is known.

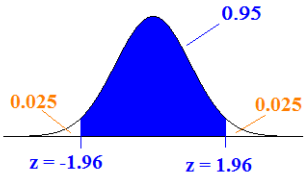
Caution

Note: The assumption that the population standard deviation is known is not usually realistic, however, we make it here to be able to introduce the concepts in the simplest case. Later, we will discuss the changes which need to be made when we do not know the population standard deviation.



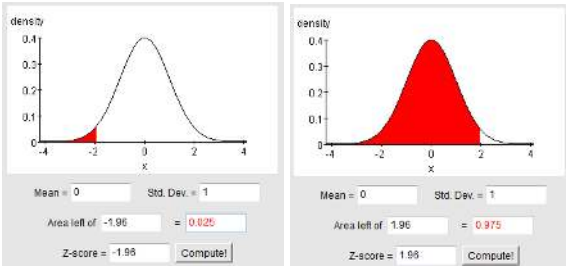
The values of x-bar follow a normal distribution with (unknown) mean μ (mu) and standard deviation σ/sqrt(n)=sigma/sqrt(n) (known, since both σ, sigma, and n are known). In the standard deviation rule, we stated that approximately 95% of values fall within 2 standard deviations of μ (mu). From now on, we will be a little more precise and use the standard normal table to find the exact value for 95%.

Our picture is as follows:



Try using the applet in the post for **Learn by Doing – Normal Random Variables** to find the cutoff illustrated above.

We can also verify the z-score using a calculator or table by finding the z-score with the area of 0.025 to the left (which would give us -1.96) or with the area to the left of 0.975 = 0.95 + 0.025 (which would give us +1.96).



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

Thus, there is a 95% chance that our sample mean x-bar will fall within 1.96*σ/sqrt(n) = 1.96*sigma/sqrt(n) of μ (mu).

Which means we are 95% confident that μ (mu) falls within 1.96*σ/sqrt(n) = 1.96*sigma/sqrt(n) of our sample mean x-bar.

Here, then, is the **general result:**

Suppose a random sample of size n is taken from a normal population of values for a quantitative variable whose mean (μ, mu) is unknown, when the standard deviation (σ, sigma) is given.

A **95% confidence interval (CI) for μ (mu)** is:

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

Comment:

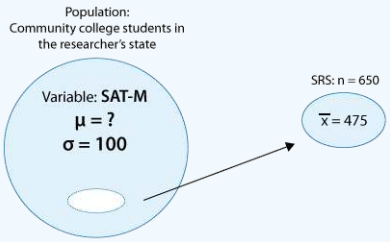
- Note that for now we require the population standard deviation (σ, sigma) to be known. Practically, σ (sigma) is rarely known, but for some cases, especially when a lot of research has been done on the quantitative variable whose mean we are estimating (such as IQ, height, weight, scores on standardized tests), it is reasonable to assume that σ (sigma) is known. Eventually, we will see how to proceed when σ (sigma) is unknown, and must be estimated with sample standard deviation (s).

Let's look at another example.

✔ EXAMPLE:

An educational researcher was interested in estimating μ (mu), the mean score on the math part of the SAT (SAT-M) of all community college students in his state. To this end, the researcher has chosen a random sample of 650 community college students from his state, and found that their average SAT-M score is 475. Based on a large body of research that was done on the SAT, it is known that the scores roughly follow a normal distribution with the standard deviation σ (sigma) =100.

Here is a visual representation of this story, which summarizes the information provided:



Based on this information, let's estimate μ (mu) with a 95% confidence interval.

Using the formula we developed earlier

$$\bar{x} \pm 1.96 * \frac{\backslash \text{simga}}{\sqrt{n}}$$

the 95% confidence interval for μ (mu) is:

$$\begin{aligned} 475 \pm 1.96 * \frac{100}{\sqrt{650}} &= \left(475 - 1.96 * \frac{100}{\sqrt{650}}, 475 + 1.96 * \frac{100}{\sqrt{650}}\right) \\ &= (475 - 7.7, 475 + 7.7) \\ &= (467.3, 482.7) \end{aligned}$$

We will usually provide information on how to round your final answer. In this case, one decimal place is enough precision for this scenario. You could also round to the nearest whole number without much loss of information here.

We are not done yet. An equally important part is to **interpret what this means in the context of the problem.**

We are 95% confident that the mean SAT-M score of all community college students in the researcher's state is covered by the interval (467.3, 482.7). Note that the confidence interval was obtained by taking 475 ± 7.7. This means that we are 95% confident that by using the sample mean (x-bar = 475) to estimate μ (mu), our error is no more than 7.7 points.

Learn by Doing: Confidence Intervals: Means #1

You just gained practice computing and interpreting a confidence interval for a population mean. Note that the way a confidence interval is used is that we hope the interval contains the population mean μ (mu). This is why we call it an "interval **for the population mean**."

The following activity is designed to help give you a better understanding of the underlying **reasoning** behind the interpretation of confidence intervals. In particular, you will gain a deeper understanding of why we say that we are "**95% confident** that the population mean is **covered** by the interval."

Learn by Doing: Connection between Confidence Intervals and Sampling Distributions with Video (1:18)

We just saw that one interpretation of a 95% confidence interval is that we are 95% confident that the population mean (μ, mu) is contained in the interval. Another useful interpretation in practice is that, given the data, the confidence interval represents the set of plausible values for the population mean μ (mu).

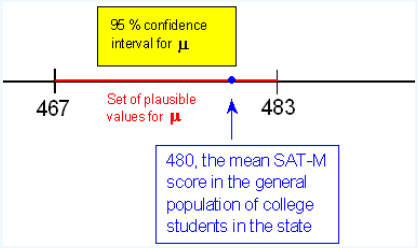
EXAMPLE:

As an illustration, let's return to the example of mean SAT-Math score of community college students. Recall that we had constructed the confidence interval (467.3, 482.7) for the unknown mean SAT-M score for all community college students.

Here is a way that we can use the confidence interval:

Do the results of this study provide evidence that μ (mu), the mean SAT-M score of community college students, is lower than the mean SAT-M score in the general population of college students in that state (which is 480)?

The 95% confidence interval for μ (mu) was found to be (467.3, 482.7). Note that 480, the mean SAT-M score in the general population of college students in that state, falls inside the interval, which means that it is one of the plausible values for μ (mu).



This means that μ (mu) could be 480 (or even higher, up to 483), and therefore we cannot conclude that the mean SAT-M score among community college students in the state is lower than the mean in the general population of college students in that state. (Note that the fact that most of the plausible values for μ (mu) fall below 480 is not a consideration here.)

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

the 95% confidence interval for μ (mu) is:

$$\begin{aligned} 475 \pm 1.96 * \frac{100}{\sqrt{650}} &= \left(475 - 1.96 * \frac{100}{\sqrt{650}}, 475 + 1.96 * \frac{100}{\sqrt{650}}\right) \\ &= (475 - 7.7, 475 + 7.7) \\ &= (467.3, 482.7) \end{aligned}$$

We will usually provide information on how to round your final answer. In this case, one decimal place is enough precision for this scenario. You could also round to the nearest whole number without much loss of information here.

We are not done yet. An equally important part is to **interpret what this means in the context of the problem.**

We are 95% confident that the mean SAT-M score of all community college students in the researcher's state is covered by the interval (467.3, 482.7). Note that the confidence interval was obtained by taking 475 ± 7.7. This means that we are 95% confident that by using the sample mean (x-bar = 475) to estimate μ (mu), our error is no more than 7.7 points.

Population Means (Part 2)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.30: Interpret confidence intervals for population parameters in context.

Learning Objectives

LO 4.31: Find confidence intervals for the population mean using the normal distribution (Z) based confidence interval formula (when required conditions are met) and perform sample size calculations.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.24: Explain the connection between the sampling distribution of a statistic, and its properties as a point estimator.

Learning Objectives

LO 6.25: Explain what a confidence interval represents and determine how changes in sample size and confidence level affect the precision of the confidence interval.

Video

Video: Population Means – Part 2 (4:04)

Other Levels of Confidence

95% is the most commonly used level of confidence. However, we may wish to increase our level of confidence and produce an interval that's almost certain to contain μ (mu). Specifically, we may want to report an interval for which we are 99% confident that it contains the unknown population mean, rather than only 95%.

Using the same reasoning as in the last comment, in order to create a 99% confidence interval for μ (mu), we should ask: There is a probability of 0.99 that any normal random variable takes values within how many standard deviations of its mean? The precise answer is 2.576, and therefore, a 99% confidence interval for μ (mu) is:

$$\bar{x} \pm 2.576 * \frac{\sigma}{\sqrt{n}}$$

Another commonly used level of confidence is a 90% level of confidence. Since there is a probability of 0.90 that any normal random variable takes values within 1.645 standard deviations of its mean, the 90% confidence interval for μ (mu) is:

$$\bar{x} \pm 1.645 * \frac{\sigma}{\sqrt{n}}$$

EXAMPLE:

Let's go back to our first example, the IQ example:

The IQ level of students at a particular university has an unknown mean (μ, mu) and known standard deviation σ (sigma) =15. A simple random sample of 100 students is found to have a sample mean IQ of 115 (x-bar). Estimate μ (mu) with a 90%, 95%, and 99% confidence interval.

A 90% confidence interval for μ (mu) is:

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}} = 115 \pm 1.645(\frac{15}{\sqrt{100}}) = 115 \pm 2.5 = (112.5, 117.5).$$

A 95% confidence interval for μ (mu) is:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 115 \pm 1.96(\frac{15}{\sqrt{100}}) = 115 \pm 2.9 = (112.1, 117.9).$$

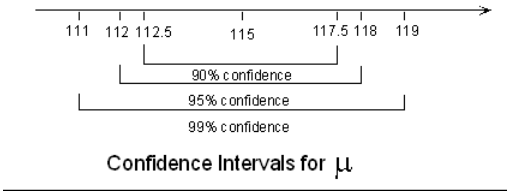
A 99% confidence interval for μ (mu) is:

$$\bar{x} \pm 2.576 \frac{\sigma}{\sqrt{n}} = 115 \pm 2.576(\frac{15}{\sqrt{100}} = 115 \pm 4.0 = (111, 119).$$

The purpose of this next activity is to give you guided practice at calculating and interpreting confidence intervals, and drawing conclusions from them.

Did I Get This?: Confidence Intervals: Means #1

Note from the previous example and the previous "Did I Get This?" activity, that the more confidence I require, the wider the confidence interval for μ (mu). The 99% confidence interval is wider than the 95% confidence interval, which is wider than the 90% confidence interval.



This is not very surprising, given that in the 99% interval we multiply the standard deviation of the statistic by 2.576, in the 95% by 2, and in the 90% only by 1.645. Beyond this numerical explanation, there is a very clear intuitive explanation and an important implication of this result.

Let's start with the intuitive explanation. The more certain I want to be that the interval contains the value of μ (mu), the more plausible values the interval needs to include in order to account for that extra certainty. I am 95% certain that the value of μ (mu) is one of the values in the interval (112.1, 117.9). In order to be 99% certain that one of the values in the interval is the value of μ (mu), I need to include more values, and thus provide a wider confidence interval.

Learn by Doing: Visualizing the Relationship between Confidence and Width

In our example, the **wider** 99% confidence interval (111, 119) gives us a **less precise** estimation about the value of μ (mu) than the narrower 90% confidence interval (112.5, 117.5), because the smaller interval 'narrows-in' on the plausible values of μ (mu).

The important practical implication here is that researchers must decide whether they prefer to state their results with a higher level of confidence or produce a more precise interval. In other words,

Caution

There is a trade-off between the level of confidence and the precision with which the parameter is estimated.

The price we have to pay for a higher level of confidence is that the unknown population mean will be estimated with less precision (i.e., with a wider confidence interval). If we would like to estimate μ (mu) with more precision (i.e. a narrower confidence interval), we will need to sacrifice and report an interval with a lower level of confidence.

Did I Get This?: Confidence Intervals: Means #2

So far we've developed the confidence interval for the population mean "from scratch" based on results from probability, and discussed the trade-off between the level of confidence and the precision of the interval. The price you pay for a higher level of confidence is a lower level of precision of the interval (i.e., a wider interval).

Is there a way to bypass this trade-off? In other words, is there a way to increase the precision of the interval (i.e., make it narrower) **without** compromising on the level of confidence? We will answer this question shortly, but first we'll need to get a deeper understanding of the different components of the confidence interval and its structure.

Understanding the General Structure of Confidence Intervals

We explored the confidence interval for μ (mu) for different levels of confidence, and found that in general, it has the following form:

$$\bar{x} \pm z^* \cdot \dfrac{\sigma}{\sqrt{n}}$$

where z* is a general notation for the multiplier that depends on the level of confidence. As we discussed before:

- For a 90% level of confidence, z* = 1.645
- For a 95% level of confidence, z* = 1.96
- For a 99% level of confidence, z* = 2.576

To start our discussion about the structure of the confidence interval, let's denote

$$m = z^* \cdot \dfrac{\sigma}{\sqrt{n}}$$

The confidence interval, then, has the form:

$$\bar{x} \pm m$$

To summarize, we have

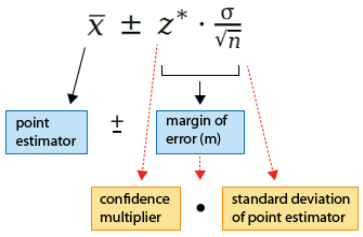
$$\bar{x} \pm \underbrace{z^* \cdot \frac{\sigma}{\sqrt{n}}}_{m}$$

X-bar is the sample mean, the point estimator for the unknown population mean (μ, mu).

m is called the **margin of error**, since it represents the maximum estimation error for a given level of confidence.

For example, for a 95% confidence interval, we are 95% confident that our estimate will not depart from the true population mean by more than m, the margin of error and m is further made up of the product of two components:

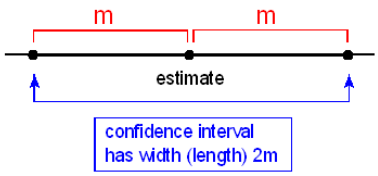
Here is a summary of the different components of the confidence interval and its structure:



This structure: **estimate ± margin of error**, where the margin of error is further composed of the product of a confidence multiplier and the standard deviation of the statistic (or, as we'll see, the standard error) is the general structure of all confidence intervals that we will encounter in this course.

Obviously, even though each confidence interval has the same components, the formula for these components is different from confidence interval to confidence interval, depending on what unknown parameter the confidence interval aims to estimate.

Since the structure of the confidence interval is such that it has a margin of error on either side of the estimate, it is centered at the estimate (in our current case, x-bar), and its width (or length) is exactly twice the margin of error:



The margin of error, m, is therefore "in charge" of the width (or precision) of the confidence interval, and the estimate is in charge of its location (and has no effect on the width).

Did I Get This?: Margin of Error

Let us now go back to the confidence interval for the mean, and more specifically, to the question that we posed at the beginning of the previous page:

Is there a way to increase the precision of the confidence interval (i.e., make it narrower) **without** compromising on the level of confidence?

Since the width of the confidence interval is a function of its margin of error, let's look closely at the margin of error of the confidence interval for the mean and see how it can be reduced:

$$m = z^* \dot \dfrac{\sigma}{\sqrt{n}}$$

Since z* controls the level of confidence, we can rephrase our question above in the following way:

Is there a way to reduce this margin of error other than by reducing z*?

If you look closely at the margin of error, you'll see that the answer is **yes.** We can do that by increasing the sample size n (since it appears in the denominator).

Many Students Wonder: **Confidence Intervals (Population Mean)**

Question*:* Isn't it true that another way to reduce the margin of error (for a fixed z*) is to reduce σ (sigma)?

Answer: While it is true that strictly mathematically speaking the smaller the value of σ (sigma), the smaller the margin of error, practically speaking we have absolutely no control over the value of σ (sigma) (i.e., we cannot make it larger or smaller). σ (sigma) is the population standard deviation; it is a fixed value (which here we assume is known) that has an effect on the width of the confidence interval (since it appears in the margin of error), but is definitely not a value we can change.

Let's look at an example first and then explain why increasing the sample size is a way to increase the precision of the confidence interval **without** compromising on the level of confidence.

✔ EXAMPLE:

Recall the IQ example:

The IQ level of students at a particular university has an unknown mean (μ, mu) and a known standard deviation of σ (sigma) =15. A simple random sample of 100 students is found to have the sample mean IQ of 115 (x-bar).

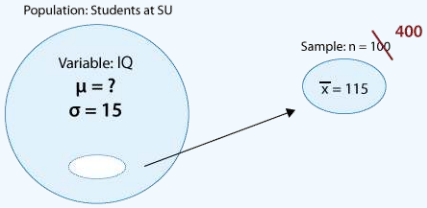
For simplicity, in this question, we will round z* = 1.96 to 2. You should use z* = 1.96 in all problems unless you are specifically instructed to do otherwise.

A 95% confidence interval for μ (mu) in this case is:

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}} = 115 \pm 2\left(\frac{15}{\sqrt{100}}\right) = 115 \pm 3.0 = (112, 118)$$

Note that the margin of error is m = 3, and therefore the width of the confidence interval is 6.

Now, what if we change the problem slightly by increasing the sample size, and assume that it was 400 instead of 100?



In this case, a 95% confidence interval for μ (mu) is:

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}} = 115 \pm 2\left(\frac{15}{\sqrt{400}}\right) = 115 \pm 1.5 = (113.5, 116.5)$$

The margin of error here is only m = 1.5, and thus the width is only 3.

Note that for the same level of confidence (95%) we now have a narrower, and thus more precise, confidence interval.

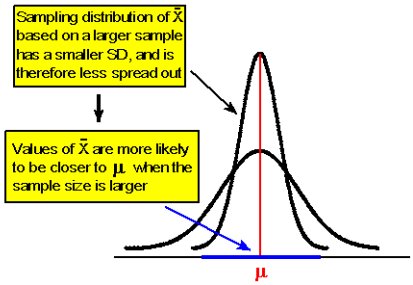
Let's try to understand why is it that a larger sample size will reduce the margin of error for a fixed level of confidence. There are three ways to explain this: mathematically, using probability theory, and intuitively.

We've already alluded to the **mathematical** explanation; the margin of error is

$$m = z^* \cdot \frac{\sigma}{\sqrt{n}}$$

and since n, the sample size, appears in the denominator, increasing n will reduce the margin of error.

As we saw in our discussion about point estimates, **probability theory** tells us that:



This explains why with a larger sample size the margin of error (which represents how far apart we believe x-bar might be from μ (mu) for a given level of confidence) is smaller.

On an intuitive level, if our estimate x-bar is based on a larger sample (i.e., a larger fraction of the population), we have more faith in it, or it is more reliable, and therefore we need to account for less error around it.

Comment:

- While it is true that for a given level of confidence, increasing the sample size increases the precision of our interval estimation, in practice, increasing the sample size is not always possible.
  - Consider a study in which there is a non-negligible cost involved for collecting data from each participant (an expensive medical procedure, for example). If the study has some budgetary constraints, which is usually the case, increasing the sample size from 100 to 400 is just not possible in terms of cost-effectiveness.
  - Another instance in which increasing the sample size is impossible is when a larger sample is simply not available, even if we had the money to afford it. For example, consider a study on the effectiveness of a drug on curing a very rare disease among children. Since the disease is rare, there are a limited number of children who could be participants.
- This is the reality of statistics. Sometimes theory collides with reality, and you simply do the best you can.

Did I Get This?: Sample Size and Confidence

Population Means (Part 3)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.30: Interpret confidence intervals for population parameters in context.

Learning Objectives

LO 4.31: Find confidence intervals for the population mean using the normal distribution (Z) based confidence interval formula (when required conditions are met) and perform sample size calculations.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.24: Explain the connection between the sampling distribution of a statistic, and its properties as a point estimator.

Learning Objectives

LO 6.25: Explain what a confidence interval represents and determine how changes in sample size and confidence level affect the precision of the confidence interval.

Video

Video: Population Means – Part 3 (6:02)

Sample Size Calculations

As we just learned, for a given level of confidence, the sample size determines the size of the margin of error and thus the width, or precision, of our interval estimation. This process can be reversed.

In situations where a researcher has some flexibility as to the sample size, the researcher can calculate in advance what the sample size is that he/she needs in order to be able to report a confidence interval with a certain level of confidence and a certain margin of error. Let's look at an example.

EXAMPLE:

Recall the example about the SAT-M scores of community college students.

An educational researcher is interested in estimating μ (mu), the mean score on the math part of the SAT (SAT-M) of all community college students in his state. To this end, the researcher has chosen a random sample of 650 community college students from his state, and found that their average SAT-M score is 475. Based on a large body of research that was done on the SAT, it is known that the scores roughly follow a normal distribution, with the standard deviation σ (sigma) =100.

The 95% confidence interval for μ (mu) is

$$\begin{aligned} 475 \pm 1.96 * \frac{100}{\sqrt{650}} &= \left(475 - 1.96 * \frac{100}{\sqrt{650}}, 475 + 1.96 * \frac{100}{\sqrt{650}}\right) \\ &= (475 - 7.7, 475 + 7.7) \\ &= (467.3, 482.7) \end{aligned}$$

which is roughly 475 ± 8, or (467, 483). For a sample size of n = 650, our margin of error is 8.

Now, let's think about this problem in a slightly different way:

An educational researcher is interested in estimating μ (mu), the mean score on the math part of the SAT (SAT-M) of all community college students in his state with a margin of error of (only) 5, at the 95% confidence level. What is the sample size needed to achieve this? σ (sigma), of course, is still assumed to be 100.

To solve this, we set:

$$m = 2 \cdot \frac{100}{\sqrt{n}} = 5 \quad \text{so} \quad \sqrt{n} = \frac{2(100)}{5} \quad \text{and} \quad n = \left(\frac{2(100)}{5}\right)^2 = 1600$$

So, for a sample size of 1,600 community college students, the researcher will be able to estimate μ (mu) with a margin of error of 5, at the 95% level. In this example, we can also imagine that the researcher has some flexibility in choosing the sample size, since there is a minimal cost (if any) involved in recording students' SAT-M scores, and there are many more than 1,600 community college students in each state.

Rather than take the same steps to isolate n every time we solve such a problem, we may obtain a general expression for the required n for a desired margin of error m and a certain level of confidence.

Since

$$\boxed{\text{m = z* \dot \dfrac{\sigma}{\sqrt{n}}}}$$

is the formula to determine m for a given n, we can use simple algebra to express n in terms of m (multiply both sides by the square root of n, divide both sides by m, and square both sides) to get

$$n = (\frac{z * \sigma}{m})^2$$

Comment:

- Clearly, the **sample size n must be an integer**.
- In the previous example we got n = 1,600, but in other situations, the calculation may give us a non-integer result.
- In these cases, we should always **round up to the next highest integer.**
- Using this "conservative approach," we'll achieve an interval at least as narrow as the one desired.

✔ EXAMPLE:

IQ scores are known to vary normally with a standard deviation of 15. How many students should be sampled if we want to estimate the population mean IQ at 99% confidence with a margin of error equal to 2?

$$n = \left(\frac{z^* \sigma}{m}\right)^2 = \left(\frac{2.576(15)}{2}\right)^2 = 373.26$$

Round up to be safe, and take a sample of 374 students.

The purpose of the next activity is to give you guided practice in sample size calculations for obtaining confidence intervals with a desired margin of error, at a certain confidence level. Consider the example from the previous Learn By Doing activity:

Learn by Doing: Sample Size

Comment:

- In the preceding activity, you saw that in order to calculate the sample size when planning a study, you needed to know the population standard deviation, sigma (σ). In practice, sigma is usually not known, because it is a parameter. (The rare exceptions are certain variables like IQ score or standardized tests that might be constructed to have a particular known sigma.)

Therefore, when researchers wish to compute the required sample size in preparation for a study, they use an **estimate** of sigma. Usually, sigma is estimated based on the standard deviation obtained in prior studies.

However, in some cases, there might not be any prior studies on the topic. In such instances, a researcher still needs to get a rough estimate of the standard deviation of the (yet-to-be-measured) variable, in order to determine the required sample size for the study. One way to get such a rough estimate is with the "range rule of thumb." We will not cover this topic in depth but mention here that a very rough estimate of the standard deviation of a population is the range/4.

There are a few more things we need to discuss:

- Is it always OK to use the confidence interval we developed for μ (mu) when σ (sigma) is known?
- What if σ (sigma) is unknown?
- How can we use statistical software to calculate confidence intervals for us?

When is it safe to use the confidence interval we developed?

One of the most important things to learn with any inference method is the conditions under which it is safe to use it. It is very tempting to apply a certain method, but if the conditions under which this method was developed are not met, then using this method will lead to unreliable results, which can then lead to wrong and/or misleading conclusions. As you'll see throughout this section, we will always discuss the conditions under which each method can be safely used.

In particular, the confidence interval for μ (mu), when σ (sigma) is known:

$$\bar{x} \pm z^* \cdot \dfrac{\sigma}{\sqrt{n}}$$

was developed assuming that the sampling distribution of x-bar is normal; in other words, that the Central Limit Theorem applies. In particular, this allowed us to determine the values of z*, the confidence multiplier, for different levels of confidence.

First, **the sample must be random.** Assuming that the sample is random, recall from the Probability unit that the Central Limit Theorem works when the **sample size is large** (a common rule of thumb for "large" is n > 30), or, for **smaller sample sizes**, if it is known that the quantitative **variable** of interest is **distributed normally** in the population. The only situation when we cannot use the confidence interval, then, is when the sample size is small and the variable of interest is not known to have a normal distribution. In that case, other methods, called non-parametric methods, which are beyond the scope of this course, need to be used. This can be summarized in the following table:

	Small sample size	Large sample size
Variable varies normally	✓	✓
Variable doesn't vary normally	✗	✓

Did I Get This?: When to Use Z-Interval (Means)

In the following activity, you have to opportunity to use software to summarize the raw data provided.

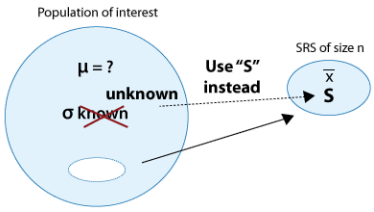
Did I Get This?: Confidence Intervals: Means #3

What if σ (sigma) is unknown?

As we discussed earlier, when variables have been well-researched in different populations it is reasonable to assume that the population standard deviation (σ, sigma) is known. However, this is rarely the case. What if σ (sigma) is unknown?

Well, there is some good news and some bad news.

The good news is that we can easily replace the population standard deviation, σ (sigma), with the **sample** standard deviation, s.



The bad news is that once σ (sigma) has been replaced by s, we lose the Central Limit Theorem, together with the normality of x-bar, and therefore the confidence multipliers z* for the different levels of confidence (1.645, 1.96, 2.576) are (generally) not correct any more. The new multipliers come from a different distribution called the "t distribution" and are therefore denoted by t* (instead of z*). We will discuss the t distribution in more detail when we talk about hypothesis testing.

The confidence interval for the population mean (μ, mu) when (σ, sigma) is unknown is therefore:

$$\bar{x} \pm t^* * \frac{s}{\sqrt{n}}$$

(Note that this interval is very similar to the one when σ (sigma) is known, with the obvious changes: s replaces σ (sigma), and t* replaces z* as discussed above.)

There is an important difference between the confidence multipliers we have used so far (z*) and those needed for the case when σ (sigma) is unknown (t*). Unlike the confidence multipliers we have used so far (z*), which depend only on the level of confidence, the new multipliers (t*) have the **added complexity** that they **depend on both the level of confidence and on the sample size** (for example: the t* used in a 95% confidence when n = 10 is different from the t* used when n = 40). Due to this added complexity in determining the appropriate t*, we will rely heavily on software in this case.

Comments:

- Since it is quite rare that σ (sigma) is known, this interval (sometimes called a "one-sample t confidence interval") is more commonly used as the confidence interval for estimating μ (mu). (Nevertheless, we could not have presented it without our extended discussion up to this point, which also provided you with a solid understanding of confidence intervals.)
- The quantity s/sqrt(n) is called the **estimated standard error** of x-bar. The Central Limit Theorem tells us that σ/sqrt(n) = sigma/sqrt(n) is the **standard deviation** of x-bar (and this is the quantity used in confidence interval when σ (sigma) is known). In general, the **standard error** is the **standard deviation of the sampling distribution of a statistic**. When we substitute s for σ (sigma) we are estimating the true standard error. You may see the term "standard error" used for both the true standard error and the estimated standard error depending on the author and audience. What is important to understand about the standard error is that it measures the variation of a statistic calculated from a sample of a specified sample size (not the variation of the original population).
- As before, to safely use this confidence interval (one-sample t confidence interval), the sample **must be random**, and the only case when this interval cannot be used is when the sample size is small and the variable is not known to vary normally.

Final Comment:

- It turns out that for large values of n, the t* multipliers are not that different from the z* multipliers, and therefore using the interval formula:

$$\bar{x} \pm z * * \frac{s}{\sqrt{n}}$$

for μ (mu) when σ (sigma) is unknown provides a pretty good approximation.

Population Means (Summary)

Let's summarize

- When the population is normal and/or the sample is large, a confidence interval for unknown population mean μ (mu) when σ (sigma) is known is:

$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

where z* is 1.645 for 90% confidence, 1.96 for 95% confidence, and 2.576 for 99% confidence.

- There is a trade-off between the level of confidence and the precision of the interval estimation. For a given sample size, the price we have to pay for more precision is sacrificing level of confidence.
- The general form of confidence intervals is an estimate +/- the margin of error (m). In this case, the estimate = x-bar and

$$\boxed{\text{m = z* \dot \dfrac{\sigma}{\sqrt{n}}}}$$

The confidence interval is therefore centered at the estimate and its width is exactly 2m.

- For a given level of confidence, the width of the interval depends on the sample size. We can therefore do a sample size calculation to figure out what sample size is needed in order to get a confidence interval with a desired margin of error m, and a certain level of confidence (assuming we have some flexibility with the sample size). To do the sample size calculation we use:

$$n = (\frac{z * \sigma}{m})^2$$

(and round **up** to the next integer). We estimate σ (sigma) when necessary.

- When σ (sigma) is unknown, we use the sample standard deviation, s, instead, but as a result we also need to use a different set of confidence multipliers (t*) associated with the t distribution. We will use software to calculate intervals in this case, however, the formula for confidence interval in this case is

$$\bar{x} \pm t * * \frac{s}{\sqrt{n}}$$

- These new multipliers have the added complexity that they depend not only on the level of confidence, but also on the sample size. Software is therefore very useful for calculating confidence intervals in this case.
- For large values of n, the t* multipliers are not that different from the z* multipliers, and therefore using the interval formula:

$$\bar{x} \pm z * * \frac{s}{\sqrt{n}}$$

for μ (mu) when σ (sigma) is unknown provides a pretty good approximation.

Population Proportions

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.30: Interpret confidence intervals for population parameters in context.

Learning Objectives

LO 4.32: Find confidence intervals for the population proportion using the formula (when required conditions are met) and perform sample size calculations.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.24: Explain the connection between the sampling distribution of a statistic, and its properties as a point estimator.

Learning Objectives

LO 6.25: Explain what a confidence interval represents and determine how changes in sample size and confidence level affect the precision of the confidence interval.

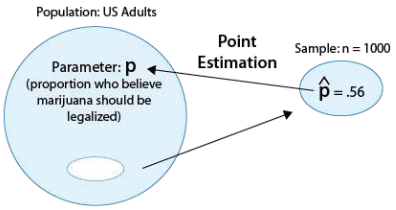
Video

Video: Population Proportions (4:13)

Confidence Intervals

As we mentioned in the introduction to Unit 4A, when the variable that we're interested in studying in the population is **categorical**, the parameter we are trying to infer about is the **population proportion (p)** associated with that variable. We also learned that the point estimator for the population proportion p is the sample proportion p-hat.

To refresh your memory, here is a picture that summarizes an example we looked at.



We are now moving on to interval estimation of p. In other words, we would like to develop a set of intervals that, with different levels of confidence, will capture the value of p. We've actually done all the groundwork and discussed all the big ideas of interval estimation when we talked about interval estimation for μ (mu), so we'll be able to go through it much faster. Let's begin.

Recall that the general form of any confidence interval for an unknown parameter is:

$$\textit{estimate} \pm \textit{margin of error}$$

Since the unknown parameter here is the population proportion p, the point estimator (as I reminded you above) is the sample proportion p-hat. The confidence interval for p, therefore, has the form:

$$\hat{p} \pm m$$

(Recall that m is the notation for the margin of error.) The margin of error (m) gives us the maximum estimation error with a certain confidence. In this case it tells us that p-hat is different from p (the parameter it estimates) by no more than m units.

From our previous discussion on confidence intervals, we also know that the margin of error is the product of two components:

$$m = \textit{confidence multiplier} \cdot \textit{SD of the estimator}$$

To figure out what these two components are, we need to go back to a result we obtained in the Sampling Distributions section of the Probability unit about the sampling distribution of p-hat. We found that under certain conditions (which we'll come back to later), p-hat has a normal distribution with mean p, and a

$$\text{standard deviation of } \sqrt{\frac{p(1-p)}{n}}$$

This result makes things very simple for us, because it reveals what the two components are that the margin of error is made of:

- Since, like the sampling distribution of x-bar, the sampling distribution of p-hat is normal, the confidence multipliers that we'll use in the confidence interval for p will be the same z* multipliers we use for the confidence interval for μ (mu) when σ (sigma) is known (using **exactly** the same reasoning and the same probability results). The multipliers we'll use, then, are: **1.645, 1.96, and 2.576 at the 90%, 95% and 99% confidence levels, respectively.**
- The standard deviation of our estimator p-hat is

$$\sqrt{\frac{p(1-p)}{n}}$$

Putting it all together, we find that the confidence interval for p should be:

$$\hat{p} \pm z^* \cdot \sqrt{\frac{p(1-p)}{n}}$$

We just have to solve one practical problem and we're done. We're trying to estimate the **unknown** population proportion *p*, so having it appear in the confidence interval doesn't make any sense. To overcome this problem, we'll do the obvious thing …

We'll replace p with its sample counterpart, p-hat, and work with the **estimated standard error of p-hat**

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Now we're done. The **confidence interval for the population proportion p** is:

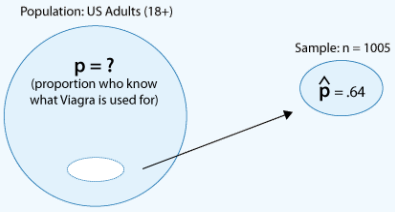
$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

EXAMPLE:

The drug Viagra became available in the U.S. in May, 1998, in the wake of an advertising campaign that was unprecedented in scope and intensity. A Gallup poll found that by the end of the first week in May, 643 out of a random sample of 1,005 adults were aware that Viagra was an impotency medication (based on "Viagra A Popular Hit," a Gallup poll analysis by Lydia Saad, May 1998).

Let's estimate the proportion p of all adults in the U.S. who by the end of the first week of May 1998 were already aware of Viagra and its purpose by setting up a 95% confidence interval for p.

We first need to calculate the sample proportion p-hat. Out of 1,005 sampled adults, 643 knew what Viagra is used for, so p-hat = 643/1005 = 0.64



Therefore, a 95% confidence interval for p is

$$\begin{aligned}\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.64 \pm 1.96 \cdot \sqrt{\frac{0.64(1-0.64)}{1005}} \\ &= 0.64 \pm 0.03 \\ &= (0.61, 0.67)\end{aligned}$$

We can be 95% confident that the proportion of all U.S. adults who were already familiar with Viagra by that time was between 0.61 and 0.67 (or 61% and 67%).

The fact that the margin of error equals 0.03 says we can be 95% confident that unknown population proportion p is within 0.03 (3%) of the observed sample proportion 0.64 (64%). In other words, we are 95% confident that 64% is "off" by no more than 3%.

Did I Get This?: Confidence Intervals – Proportions #1

Comment:

- We would like to share with you the methodology portion of the official poll release for the Viagra example. We hope you see that you now have the tools to understand how poll results are analyzed:

"The results are based on telephone interviews with a randomly selected national sample of 1,005 adults, 18 years and older, conducted May 8-10, 1998. For results based on samples of this size, one can say with 95 percent confidence that the error attributable to sampling and other random effects could be plus or minus 3 percentage points. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls."

The purpose of the next activity is to provide guided practice in calculating and interpreting the confidence interval for the population proportion p, and drawing conclusions from it.

> **Learn by Doing:** Confidence Intervals – Proportions #1

Two important results that we discussed at length when we talked about the confidence interval for μ (mu) also apply here:

1. There is a trade-off between level of confidence and the width (or precision) of the confidence interval. The more precision you would like the confidence interval for p to have, the more you have to pay by having a lower level of confidence.

2. Since n appears in the denominator of the margin of error of the confidence interval for p, for a fixed level of confidence, the larger the sample, the narrower, or more precise it is. This brings us naturally to our next point.

Sample Size Calculations

Just as we did for means, when we have some level of flexibility in determining the sample size, we can set a desired margin of error for estimating the population proportion and find the sample size that will achieve that.

For example, a final poll on the day before an election would want the margin of error to be quite small (with a high level of confidence) in order to be able to predict the election results with the most precision. This is particularly relevant when it is a close race between the candidates. The polling company needs to figure out how many eligible voters it needs to include in their sample in order to achieve that.

Let's see how we do that.

(**Comment:** For our discussion here we will focus on a 95% confidence level ($z^* = 1.96$), since this is the most commonly used level of confidence.)

The confidence interval for p is

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The margin of error, then, is

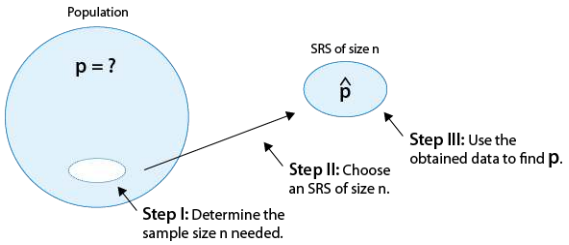
$$m = 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Now we isolate n (i.e., express it as a function of m).

$$n = \frac{(1.96)^2 \hat{p}(1-\hat{p})}{m^2}$$

There is a practical problem with this expression that we need to overcome.

Practically, you first determine the sample size, then you choose a random sample of that size, and then use the collected data to find p-hat.



So the fact that the expression above for determining the sample size depends on p-hat is problematic.

The way to overcome this problem is to take the conservative approach by setting p-hat = 1/2 = 0.5.

Why do we call this approach conservative?

It is conservative because the expression that appears in the numerator,

$$\hat{p}(1-\hat{p})$$

is maximized when p-hat = 1/2 = 0.5.

That way, the n we get will work in giving us the desired margin of error regardless of what the value of p-hat is. This is a "worst case scenario" approach. So when we do that we get:

$$n = \frac{(1.96)^2 \frac{1}{2}(1-\frac{1}{2})}{m^2} = \frac{(1.96)^2}{4 \cdot m^2}$$

In general, for any confidence level we have

- If we know a reasonable estimate of the proportion we can use:

$$n = \frac{(z^*)^2 \hat{p}(1-\hat{p})}{m^2}$$

- If we choose the conservative estimate assuming we know nothing about the true proportion we use:

$$n = \frac{(z^*)^2}{4 \cdot m^2}$$

EXAMPLE:

It seems like media polls usually use a sample size of 1,000 to 1,200. This could be puzzling.

How could the results obtained from, say, 1,100 U.S. adults give us information about the entire population of U.S. adults? 1,100 is such a tiny fraction of the actual population. Here is the answer:

What sample size n is needed if a margin of error m = 0.03 is desired?

$$n = \frac{(1.96)^2}{4 \cdot (0.03)^2} = 1067.1 \rightarrow 1068$$

(remember, always round up). In fact, 0.03 is a very commonly used margin of error, especially for media polls. For this reason, most media polls work with a sample of around 1,100 people.

Did I Get This?: Confidence Intervals – Proportions #2

When is it safe to use these methods?

As we mentioned before, one of the most important things to learn with any inference method is the conditions under which it is safe to use it.

As we did for the mean, the assumption we made in order to develop the methods in this unit was that the sampling distribution of the sample proportion, p-hat is roughly normal. Recall from the Probability unit that the conditions under which this happens are

that

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

> Since p is unknown, we will replace it with its estimate, the sample proportion, and set
>
> $$n\hat{p} \geq 10 \text{ and } n(1-\hat{p}) \geq 10$$
>
> to be the conditions under which it is safe to use the methods we developed in this section.

Here is one final practice for these confidence intervals!!

> **Did I Get This?:** Confidence Intervals – Proportions #3

Let's summarize

In general, a confidence interval for the unknown population proportion (p) is

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z* is 1.645 for 90% confidence, 1.96 for 95% confidence, and 2.576 for 99% confidence.

To obtain a desired margin of error (m) in a confidence interval for an unknown population proportion, a conservative sample size is

$$n = \frac{(z^*)^2}{4 \cdot m^2}$$

If a reasonable estimate of the true proportion is known, the sample size can be calculated using

$$n = \frac{(1.96)^2 \hat{p}(1-\hat{p})}{m^2}$$

The methods developed in this unit are safe to use as long as

$$n\hat{p} \geq 10 \text{ and } n(1-\hat{p}) \geq 10$$

Wrap-Up (Estimation)

In this section on estimation, we have discussed the basic process for constructing confidence intervals from point estimates. In doing so we must calculate the margin of error using the standard error (or estimated standard error) and a z* or t* value.

As we wrap up this topic, we wanted to again discuss the interpretation of a confidence interval.

What do we mean by "confidence"?

Suppose we find a 95% confidence interval for an unknown parameter, what does the 95% mean exactly?

- If we repeat the process for all possible samples of this size for the population, 95% of the intervals we construct will contain the parameter

This is NOT the same as saying "*the **probability** that μ (mu) is contained in (the interval constructed from my sample) is 95%.*" Why?!

Answer

- Once we have a particular confidence interval, the true value is either in the interval constructed from our sample (probability = 1) or it is not (probability = 0). We simply do not know which it is. If we were to say "the probability that μ (mu) is contained in (the interval constructed from my sample) is 95%," we know we would be incorrect since it is either 0 (No) or 1 (Yes) for any given sample. The probability comes from the "long run" view of the process.
- The probability we used to construct the confidence interval was based upon the fact that the sample statistic (x-bar, p-hat) will vary in a manner we understand (because we know the sampling distribution).

- The probability is associated with the randomness of our statistic so that for a particular interval we only speak of being “95% confident” which translates into an understanding about the process.
- **In other words, in statistics, “95% confident” means our confidence in the process and implies that in the long run, we will be correct by using this process 95% of the time but that 5% of the time we will be incorrect. For one particular use of this process we cannot know if we are one of the 95% which are correct or one of the 5% which are incorrect. That is the statistical definition of confidence.**
- We can say that in the long run, 95% of these intervals will contain the true parameter and 5% will not.

Correct Interpretations:

Example: Suppose a 95% confidence interval for the proportion of U.S. adults who are not active at all is (0.23, 0.27).

- **Correct Interpretation #1:** We are 95% confident that the true proportion of U.S. adults who are not active at all is between 23% and 27%
- **Correct Interpretation #2:** We are 95% confident that the true proportion of U.S. adults who are not active at all is covered by the interval (23%, 27%)
- **A More Thorough Interpretation:** Based upon our sample, the true proportion of U.S. adults who are not active at all is estimated to be 25%. With 95% confidence, this value could be as small as 23% to as large as 27%.
- **A Common Interpretation in Journal Articles:** Based upon our sample, the true proportion of U.S. adults who are not active at all is estimated to be 25% (95% CI 23%-27%).

Now let’s look at an INCORRECT interpretation which we have seen before

- **INCORRECT Interpretation:** *There is a 95% chance that the true proportion of U.S. adults who are not active at all is between 23% and 27%*. We know this is incorrect because at this point, the true proportion and the numbers in our interval are fixed. The probability is either 1 or 0 depending on whether the interval is one of the 95% that cover the true proportion, or one of the 5% that do not.

For confidence intervals regarding a population mean, we have an additional caution to discuss about interpretations.

Example: Suppose a 95% confidence interval for the average minutes per day of exercise for U.S. adults is (12, 18).

- **Correct Interpretation:** We are 95% confident that the true mean minutes per day of exercise for U.S. adults is between 12 and 18 minutes.
- **INCORRECT Interpretation:** We are 95% confident that an individual U.S. adult exercises between 12 and 18 minutes per day. We must remember that our intervals are about the parameter, in this case the population mean. They do not apply to an individual as we expect individuals to have much more variation.
- **INCORRECT Interpretation:** We are 95% confident that U.S. adults exercise between 12 and 18 minutes per day.This interpretation is implying this is true for all U.S. adults. This is an incorrect interpretation for the same reason as the previous incorrect interpretation!

As we continue to study inferential statistics, we will see that confidence intervals are used in many situations. The goal is always to provide confidence in our interval estimate of a quantity of interest. Population means and proportions are common parameters, however, any quantity that can be estimated from data has a population counterpart which we may wish to estimate.

> **(Optional) Outside Reading:** Little Handbook – Confidence Intervals (and More) (4 Readings, ≈ 5500 words)

Hypothesis Testing

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Learning Objectives

LO 6.27: Explain what the p-value is and how it is used to draw conclusions.

Video

Video: Hypothesis Testing (8:43)

Introduction

We are in the middle of the part of the course that has to do with inference for one variable.

So far, we talked about point estimation and learned how interval estimation enhances it by quantifying the magnitude of the estimation error (with a certain level of confidence) in the form of the margin of error. The result is the confidence interval — an interval that, with a certain confidence, we believe captures the unknown parameter.

We are now moving to the other kind of inference, **hypothesis testing**. We say that hypothesis testing is "the other kind" because, unlike the inferential methods we presented so far, where the goal was **estimating** the unknown parameter, the idea, logic and goal of hypothesis testing are quite different.

In the first two parts of this section we will discuss the idea behind hypothesis testing, explain how it works, and introduce new terminology that emerges in this form of inference. The final two parts will be more specific and will discuss hypothesis testing for the population proportion (p) and the population mean (μ, mu).

If this is your first statistics course, you will need to spend considerable time on this topic as there are many new ideas. Many students find this process and its logic difficult to understand in the beginning.

In this section, we will use the hypothesis test for a population proportion to motivate our understanding of the process. We will conduct these tests manually. For all future hypothesis test procedures, including problems involving means, we will use software to obtain the results and focus on interpreting them in the context of our scenario.

General Idea and Logic of Hypothesis Testing

The purpose of this section is to gradually build your understanding about how statistical hypothesis testing works. We start by explaining the general logic behind the process of hypothesis testing. Once we are confident that you understand this logic, we will add some more details and terminology.

To start our discussion about the idea behind statistical hypothesis testing, consider the following example:

EXAMPLE:

A case of suspected cheating on an exam is brought in front of the disciplinary committee at a certain university.

There are **two** opposing **claims** in this case:

- The **student's claim:** I did not cheat on the exam.
- The **instructor's claim:** The student did cheat on the exam.

Adhering to the principle **"innocent until proven guilty,"** the committee asks the instructor for **evidence** to support his claim. The instructor explains that the exam had two versions, and shows the committee members that on three separate exam questions, the student used in his solution numbers that were given in the other version of the exam.

The committee members all agree that **it would be extremely unlikely to get evidence like that if the student's claim of not cheating had been true.** In other words, the committee members all agree that the instructor brought forward strong enough evidence to reject the student's claim, and conclude that the student did cheat on the exam.

What does this example have to do with statistics?

While it is true that this story seems unrelated to statistics, it captures all the elements of hypothesis testing and the logic behind it. Before you read on to understand why, it would be useful to read the example again. Please do so now.

Statistical hypothesis testing is defined as:

- **Assessing evidence provided by the data against the null claim (the claim which is to be assumed true unless enough evidence exists to reject it).**

Here is how the process of statistical hypothesis testing works:

1. We have **two claims** about what is going on in the population. Let's call them **claim 1 (this will be the null claim or hypothesis)** and **claim 2 (this will be the alternative)**. Much like the story above, where the student's claim is challenged by the instructor's claim, the null claim 1 is challenged by the alternative claim 2. (For us, these claims are usually about the value of population parameter(s) or about the existence or nonexistence of a relationship between two variables in the population).
2. We choose a sample, collect relevant data and summarize them (this is similar to the instructor collecting evidence from the student's exam). For statistical tests, this step will also involve checking any conditions or assumptions.
3. We figure out how likely it is to observe data like the data we obtained, if claim 1 is true. (Note that the wording "how likely …" implies that this step requires some kind of probability calculation). In the story, the committee members assessed how likely it is to observe evidence such as the instructor provided, had the student's claim of not cheating been true.
4. Based on what we found in the previous step, we make our decision:
    - If, after assuming claim 1 is true, we find that it would be **extremely unlikely** to observe data as strong as ours or stronger in favor of claim 2, then we have strong evidence against claim 1, and we reject it in favor of claim 2. Later we will see this corresponds to a small p-value.
    - If, after assuming claim 1 is true, we find that observing data as strong as ours or stronger in favor of claim 2 is **NOT VERY UNLIKELY**, then we do not have enough evidence against claim 1, and therefore we cannot reject it in favor of claim 2. Later we will see this corresponds to a p-value which is not small.

In our story, the committee decided that it would be extremely unlikely to find the evidence that the instructor provided had the student's claim of not cheating been true. In other words, the members felt that it is extremely unlikely that it is just a coincidence (random chance) that the student used the numbers from the other version of the exam on three separate problems. The committee members therefore decided to reject the student's claim and concluded that the student had, indeed, cheated on the exam. (Wouldn't you conclude the same?)

Hopefully this example helped you understand the logic behind hypothesis testing.

Interactive Applet: Reasoning of a Statistical Test

To strengthen your understanding of the process of hypothesis testing and the logic behind it, let's look at three statistical examples.

✔ EXAMPLE:

A recent study estimated that 20% of all college students in the United States smoke. The head of Health Services at Goodheart University (GU) suspects that the proportion of smokers may be lower at GU. In hopes of confirming her claim, the head of Health Services chooses a random sample of 400 Goodheart students, and finds that 70 of them are smokers.

Let's analyze this example using the 4 steps outlined above:

1. **Stating the claims:** There are two claims here:
   - **claim 1:** The proportion of smokers at Goodheart is 0.20.
   - **claim 2:** The proportion of smokers at Goodheart is less than 0.20.

   Claim 1 basically says "nothing special goes on at Goodheart University; the proportion of smokers there is no different from the proportion in the entire country." This claim is challenged by the head of Health Services, who suspects that the proportion of smokers at Goodheart is lower.

2. **Choosing a sample and collecting data:** A sample of n = 400 was chosen, and summarizing the data revealed that the sample proportion of smokers is *p*-hat = 70/400 = 0.175.While it is true that 0.175 is less than 0.20, it is not clear whether this is strong enough evidence against claim 1. We must account for sampling variation.

3. **Assessment of evidence:** In order to assess whether the data provide strong enough evidence against claim 1, we need to ask ourselves: How surprising is it to get a sample proportion as low as *p*-hat = 0.175 (or lower), assuming claim 1 is true? In other words, we need to find how likely it is that in a random sample of size n = 400 taken from a population where the proportion of smokers is $p = 0.20$ we'll get a sample proportion as low as *p*-hat = 0.175 (or lower).It turns out that the probability that we'll get a sample proportion as low as *p*-hat = 0.175 (or lower) in such a sample is roughly 0.106 (do not worry about how this was calculated at this point – however, if you think about it hopefully you can see that the key is the sampling distribution of *p*-hat).

4. **Conclusion:** Well, we found that if claim 1 were true there is a probability of 0.106 of observing data like that observed or more extreme. Now you have to decide …Do you think that a probability of 0.106 makes our data rare enough (surprising enough) under claim 1 so that the fact that we **did** observe it is enough evidence to reject claim 1? Or do you feel that a probability of 0.106 means that data like we observed are not very likely when claim 1 is true, but they are not unlikely enough to conclude that getting such data is sufficient evidence to reject claim 1. Basically, this is your decision. However, it would be nice to have some kind of guideline about what is generally considered surprising enough.

✔ EXAMPLE:

A certain prescription allergy medicine is supposed to contain an average of 245 parts per million (ppm) of a certain chemical. If the concentration is higher than 245 ppm, the drug will likely cause unpleasant side effects, and if the concentration is below 245 ppm, the drug may be ineffective. The manufacturer wants to check whether the mean concentration in a large shipment is the required 245 ppm or not. To this end, a random sample of 64 portions from the large shipment is tested, and it is found that the sample mean concentration is 250 ppm with a sample standard deviation of 12 ppm.

1. **Stating the claims:**
   - **Claim 1:** The mean concentration in the shipment is the required 245 ppm.
   - **Claim 2:** The mean concentration in the shipment is not the required 245 ppm.

   Note that again, claim 1 basically says: "There is nothing unusual about this shipment, the mean concentration is the required 245 ppm." This claim is challenged by the manufacturer, who wants to check whether that is, indeed, the case or not.

2. **Choosing a sample and collecting data:** A sample of n = 64 portions is chosen and after summarizing the data it is found that the sample mean concentration is x-bar = 250 and the sample standard deviation is s = 12.Is the fact that x-bar = 250 is different from 245 strong enough evidence to reject claim 1 and conclude that the mean concentration in the whole shipment is not the required 245? In other words, do the data provide strong enough evidence to reject claim 1?

3. **Assessing the evidence:** In order to assess whether the data provide strong enough evidence against claim 1, we need to ask ourselves the following question: If the mean concentration in the whole shipment were really the required 245 ppm (i.e., if claim 1 were true), how surprising would it be to observe a sample of 64 portions where the sample mean concentration is off by 5 ppm or more (as we did)? It turns out that it would be extremely unlikely to get such a result if the mean concentration were really the required 245. There is only a probability of 0.0007 (i.e., 7 in 10,000) of that happening. (Do not worry about how this was calculated at this point, but again, the key will be the sampling distribution.)

4. **Making conclusions:** Here, it is pretty clear that a sample like the one we observed or more extreme is VERY rare (or extremely unlikely) if the mean concentration in the shipment were really the required 245 ppm. The fact that we **did** observe such a sample therefore provides strong evidence against claim 1, so we reject it and conclude with very little doubt that the mean concentration in the shipment is not the required 245 ppm.

Do you think that you're getting it? Let's make sure, and look at another example.

✔ EXAMPLE:

Is there a relationship between gender and combined scores (Math + Verbal) on the SAT exam?

Following a report on the College Board website, which showed that in 2003, males scored generally higher than females on the SAT exam, an educational researcher wanted to check whether this was also the case in her school district. The researcher chose random samples of 150 males and 150 females from her school district, collected data on their SAT performance and found the following:

Females			Males		
n	**mean**	**standard deviation**	**n**	**mean**	**standard deviation**
150	1010	206	150	1025	212

Again, let's see how the process of hypothesis testing works for this example:

1. **Stating the claims:**
   - **Claim 1:** Performance on the SAT is not related to gender (males and females score the same).
   - **Claim 2:** Performance on the SAT is related to gender – males score higher.

   Note that again, claim 1 basically says: "There is nothing going on between the variables SAT and gender." Claim 2 represents what the researcher wants to check, or suspects might actually be the case.

2. **Choosing a sample and collecting data:** Data were collected and summarized as given above. Is the fact that the sample mean score of males (1,025) is higher than the sample mean score of females (1,010) by 15 points strong enough information to reject claim 1 and conclude that in this researcher's school district, males score higher on the SAT than females?

3. **Assessment of evidence:** In order to assess whether the data provide strong enough evidence against claim 1, we need to ask ourselves: If SAT scores are in fact not related to gender (claim 1 is true), how likely is it to get data like the data we observed, in which the difference between the males' average and females' average score is as high as 15 points or higher? It turns out that the probability of observing such a sample result if SAT score is not related to gender is approximately 0.29 (Again, do not worry about how this was calculated at this point).

4. **Conclusion:** Here, we have an example where observing a sample like the one we observed or more extreme is definitely not surprising (roughly 30% chance) if claim 1 were true (i.e., if indeed there is no difference in SAT scores between males and females). We therefore conclude that our data does not provide enough evidence for rejecting claim 1.

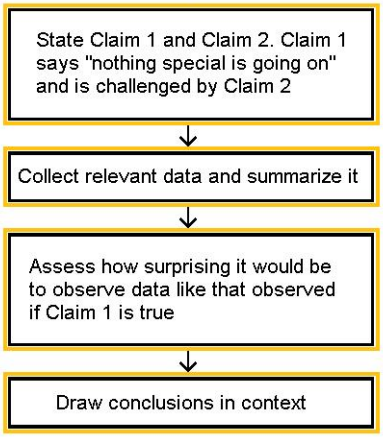
Comment:

- Go back and read the conclusion sections of the three examples, and pay attention to the wording. Note that there are two types of conclusions:
  - "The data provide enough evidence to reject claim 1 and accept claim 2"; or
  - "The data do not provide enough evidence to reject claim 1."

In particular, note that in the second type of conclusion **we did not say:** "**I accept claim 1**," but only "**I don't have enough evidence to reject claim 1**." We will come back to this issue later, but this is a good place to make you aware of this subtle

difference.

Hopefully by now, you understand the logic behind the statistical hypothesis testing process. Here is a summary:



Learn by Doing: Logic of Hypothesis Testing

Did I Get This?: Logic of Hypothesis Testing

Steps in Hypothesis Testing

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Learning Objectives

LO 6.27: Explain what the p-value is and how it is used to draw conclusions.

Video

Video: Steps in Hypothesis Testing (16:02)

Now that we understand the general idea of how statistical hypothesis testing works, let's go back to each of the steps and delve slightly deeper, getting more details and learning some terminology.

Hypothesis Testing Step 1: State the Hypotheses

In all three examples, our aim is to decide between two opposing points of view, Claim 1 and Claim 2. In hypothesis testing, **Claim 1** is called the **null hypothesis** (denoted "**Ho**"), and **Claim 2** plays the role of the **alternative hypothesis** (denoted "**Ha**"). As we saw in the three examples, the null hypothesis suggests nothing special is going on; in other words, there is no change from the status quo, no difference from the traditional state of affairs, no relationship. In contrast, the alternative hypothesis disagrees with this, stating that something is going on, or there is a change from the status quo, or there is a difference from the traditional state of affairs. The alternative hypothesis, Ha, usually represents what we want to check or what we suspect is really going on.

Let's go back to our three examples and apply the new notation:

In example 1:

- **Ho:** The proportion of smokers at GU is 0.20.
- **Ha:** The proportion of smokers at GU is less than 0.20.

In example 2:

- **Ho:** The mean concentration in the shipment is the required 245 ppm.
- **Ha:** The mean concentration in the shipment is not the required 245 ppm.

In example 3:

- **Ho:** Performance on the SAT is not related to gender (males and females score the same).
- **Ha:** Performance on the SAT is related to gender – males score higher.

> **Learn by Doing:** State the Hypotheses

> **Did I Get This?:** State the Hypotheses

Hypothesis Testing Step 2: Collect Data, Check Conditions and Summarize Data

This step is pretty obvious. This is what inference is all about. You look at sampled data in order to draw conclusions about the entire population. In the case of hypothesis testing, based on the data, you draw conclusions about whether or not there is enough evidence to reject Ho.

There is, however, one detail that we would like to add here. In this step we collect data and **summarize** it. Go back and look at the second step in our three examples. Note that in order to summarize the data we used simple sample statistics such as the sample proportion (*p*-hat), sample mean (x-bar) and the sample standard deviation (s).

In practice, you go a step further and use these sample statistics to summarize the data with what's called a **test statistic**. We are not going to go into any details right now, but we will discuss test statistics when we go through the specific tests.

This step will also involve checking any conditions or assumptions required to use the test.

Hypothesis Testing Step 3: Assess the Evidence

As we saw, this is the step where we calculate how likely is it to get data like that observed (or more extreme) when Ho is true. In a sense, this is the heart of the process, since we draw our conclusions based on this probability.

- If this probability is very small (see example 2), then that means that it would be very surprising to get data like that observed (or more extreme) if Ho were true. The fact that we **did** observe such data is therefore evidence against Ho, and we should reject it.
- On the other hand, if this probability is not very small (see example 3) this means that observing data like that observed (or more extreme) is not very surprising if Ho were true. The fact that we observed such data does not provide evidence against Ho. This crucial probability, therefore, has a special name. It is called the **p-value** of the test.

In our three examples, the p-values were given to you (and you were reassured that you didn't need to worry about how these were derived yet):

- Example 1: p-value = 0.106
- Example 2: p-value = 0.0007
- Example 3: p-value = 0.29

Obviously, the smaller the p-value, the more surprising it is to get data like ours (or more extreme) when Ho is true, and therefore, the stronger the evidence the data provide against Ho.

Looking at the three p-values of our three examples, we see that the data that we observed in example 2 provide the strongest evidence against the null hypothesis, followed by example 1, while the data in example 3 provides the least evidence against Ho.

Comment:

- Right now we will not go into specific details about p-value calculations, but just mention that since the p-value is the probability of getting **data** like those observed (or more extreme) when Ho is true, it would make sense that the calculation of

the p-value will be based on the data summary, which, as we mentioned, is the test statistic. Indeed, this is the case. In practice, we will mostly use software to provide the p-value for us.

Hypothesis Testing Step 4: Making Conclusions

Since our statistical conclusion is based on how small the p-value is, or in other words, how surprising our data are when Ho is true, it would be nice to have some kind of guideline or cutoff that will help determine how small the p-value must be, or how "rare" (unlikely) our data must be when Ho is true, for us to conclude that we have enough evidence to reject Ho.

This cutoff exists, and because it is so important, it has a special name. It is called the **significance level of the test** and is usually denoted by the Greek letter α (alpha). The most commonly used significance level is α (alpha) = 0.05 (or 5%). This means that:

- if the p-value < α (alpha) (usually 0.05), then the data we obtained is considered to be "rare (or surprising) enough" under the assumption that Ho is true, and we say that the data provide statistically significant evidence against Ho, so we reject Ho and thus accept Ha.
- if the p-value > α (alpha)(usually 0.05), then our data are not considered to be "surprising enough" under the assumption that Ho is true, and we say that our data do not provide enough evidence to reject Ho (or, equivalently, that the data do not provide enough evidence to accept Ha).

Now that we have a cutoff to use, here are the appropriate conclusions for each of our examples based upon the p-values we were given.

In Example 1:

- Using our cutoff of 0.05, we fail to reject Ho.
- **Conclusion**: There **IS NOT** enough evidence that the proportion of smokers at GU is less than 0.20
- **Still we should consider:** Does the evidence seen in the data provide any practical evidence towards our alternative hypothesis?

In Example 2:

- Using our cutoff of 0.05, we reject Ho.
- **Conclusion**: There **IS** enough evidence that the mean concentration in the shipment is not the required 245 ppm.
- **Still we should consider: Does the evidence seen in the data provide any practical evidence towards our alternative hypothesis?**

In Example 3:

- Using our cutoff of 0.05, we fail to reject Ho.
- **Conclusion**: There **IS NOT** enough evidence that males score higher on average than females on the SAT.
- **Still we should consider:** Does the evidence seen in the data provide any practical evidence towards our alternative hypothesis?

Notice that all of the above conclusions are written in terms of the alternative hypothesis and are given in the context of the situation. In no situation have we claimed the null hypothesis is true. Be very careful of this and other issues discussed in the following comments.

Comments:

1. Although the significance level provides a good guideline for drawing our conclusions, it should not be treated as an incontrovertible truth. There is a lot of room for personal interpretation. What if your p-value is 0.052? You might want to stick to the rules and say "0.052 > 0.05 and therefore I don't have enough evidence to reject Ho", but you might decide that 0.052 is small enough for you to believe that Ho should be rejected. It should be noted that scientific journals do consider 0.05 to be the cutoff point for which any p-value below the cutoff indicates enough evidence against Ho, and any p-value above it, **or even equal to it**, indicates there is not enough evidence against Ho. Although a p-value between 0.05 and 0.10 is often reported as marginally statistically significant.

---

2. It is important to draw your conclusions **in context**. It is **never enough** to say: **"p-value = …, and therefore I have enough evidence to reject Ho at the 0.05 significance level."** You **should always word your conclusion in terms of the data.** Although we will use the terminology of "rejecting Ho" or "failing to reject Ho" – this is mostly due to the fact that we are

instructing you in these concepts. In practice, this language is rarely used. We also suggest writing your conclusion in terms of the alternative hypothesis.Is there or is there not enough evidence that the alternative hypothesis is true?

---

3. Let's go back to the issue of the nature of the two types of conclusions that I can make.

- *Either* **I reject Ho (when the p-value is smaller than the significance level)**
- *or* **I cannot reject Ho (when the p-value is larger than the significance level).**

As we mentioned earlier, note that the second conclusion does not imply that I accept Ho, but just that I don't have enough evidence to reject it. Saying (by mistake) "I don't have enough evidence to reject Ho so I accept it" indicates that the data provide evidence that Ho is true, which is **not necessarily the case**. Consider the following slightly artificial yet effective example:

✔ EXAMPLE:

An employer claims to subscribe to an "equal opportunity" policy, not hiring men any more often than women for managerial positions. Is this credible? You're not sure, so you want to test the following **two hypotheses:**

- **Ho:** The proportion of male managers hired is 0.5
- **Ha:** The proportion of male managers hired is more than 0.5

Data: You choose at random three of the new managers who were hired in the last 5 years and find that all 3 are men.

Assessing Evidence: If the proportion of male managers hired is really 0.5 (Ho is true), then the probability that the random selection of three managers will yield three males is therefore 0.5 * 0.5 * 0.5 = 0.125. This is the p-value (using the multiplication rule for independent events).

Conclusion: Using 0.05 as the significance level, you conclude that since the p-value = 0.125 > 0.05, the fact that the three randomly selected managers were all males is not enough evidence to reject the employer's claim of subscribing to an equal opportunity policy (Ho).

However, **the data (all three selected are males) definitely does NOT provide evidence to accept the employer's claim (Ho).**

Learn By Doing: Using p-values

Did I Get This?: Using p-values

Comment about wording: Another common wording in scientific journals is:

- "The results are statistically significant" – when the p-value < α (alpha).
- "The results are not statistically significant" – when the p-value > α (alpha).

Often you will see significance levels reported with additional description to indicate the degree of statistical significance. A general guideline (although not required in our course) is:

- If 0.01 ≤ p-value < 0.05, then the results are (statistically) *significant*.
- If 0.001 ≤ p-value < 0.01, then the results are *highly statistically significant*.
- If p-value < 0.001, then the results are *very highly statistically significant*.
- If p-value > 0.05, then the results are *not statistically significant* (NS).
- If 0.05 ≤ p-value < 0.10, then the results are *marginally statistically significant*.

Let's summarize

We learned quite a lot about hypothesis testing. We learned the logic behind it, what the key elements are, and what types of conclusions we can and cannot draw in hypothesis testing. Here is a quick recap:

Video

Video: Hypothesis Testing Overview (2:20)

Here are a few more activities if you need some additional practice.

Did I Get This?: Hypothesis Testing Overview

Comments:

- Notice that **the p-value is an example of a conditional probability**. We calculate the probability of obtaining results like those of our data (or more extreme) GIVEN the null hypothesis is true. We could write P(Obtaining results like ours or more extreme | Ho is True).
- Another common phrase used to define the p-value is: "**The probability of obtaining a statistic as or more extreme than your result given the null hypothesis is TRUE**".
  - We could write P(Obtaining a test statistic as or more extreme than ours | Ho is True).
  - In this case we are asking "Assuming the null hypothesis is true, how rare is it to observe something as or more extreme than what I have found in my data?"
  - If after assuming the null hypothesis is true, what we have found in our data is extremely rare (small p-value), this provides evidence to reject our assumption that Ho is true in favor of Ha.
- The **p-value can also be thought of as the probability, assuming the null hypothesis is true, that the result we have seen is solely due to random error (or random chance).** We have already seen that statistics from samples collected from a population vary. There is random error or random chance involved when we sample from populations.

In this setting, if the p-value is very small, this implies, assuming the null hypothesis is true, that it is extremely unlikely that the results we have obtained would have happened due to random error alone, and thus our assumption (Ho) is rejected in favor of the alternative hypothesis (Ha).

- **It is EXTREMELY important that you find a definition of the p-value which makes sense to you. New students often need to contemplate this idea repeatedly through a variety of examples and explanations before becoming comfortable with this idea. It is one of the two most important concepts in statistics (the other being confidence intervals).**

Remember:

- We infer that the alternative hypothesis is true ONLY by rejecting the null hypothesis.
- A statistically significant result is one that has a very low probability of occurring if the null hypothesis is true.
- Results which are **statistically** significant may or may not have **practical** significance and vice versa.

Error and Power

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.28: Define a Type I and Type II error in general and in the context of specific scenarios.

Learning Objectives

LO 6.29: Explain the concept of the power of a statistical test including the relationship between power, sample size, and effect size.

Video

Video: Errors and Power (12:03)

Type I and Type II Errors in Hypothesis Tests

We have not yet discussed the fact that we are not guaranteed to make the correct decision by this process of hypothesis testing. Maybe you are beginning to see that there is always some level of uncertainty in statistics.

Let's think about what we know already and define the possible errors we can make in hypothesis testing. When we conduct a hypothesis test, we choose one of two possible conclusions based upon our data.

If the **p-value is smaller than your pre-specified significance level (α, alpha), you reject the null hypothesis** and either

- You have made the correct decision since the null hypothesis is false

OR

- You have made an error (**Type I**) and rejected Ho when in fact Ho is true (your data happened to be a RARE EVENT under Ho)

If the **p-value is greater than (or equal to) your chosen significance level (α, alpha), you fail to reject the null hypothesis** and either

- You have made the correct decision since the null hypothesis is true

OR

- You have made an error (**Type II**) and failed to reject Ho when in fact Ho is false (the alternative hypothesis, Ha, is true)

The following summarizes the four possible results which can be obtained from a hypothesis test. Notice the rows represent the decision made in the hypothesis test and the columns represent the (usually unknown) truth in reality.

Although the truth is unknown in practice – or we would not be conducting the test – we know it must be the case that either the null hypothesis is true or the null hypothesis is false. It is also the case that **either decision we make in a hypothesis test can result in an incorrect conclusion!**

A **TYPE I Error** occurs when we Reject Ho when, in fact, Ho is True. In this case, **we mistakenly reject a true null hypothesis.**

- P(TYPE I Error) = P(Reject Ho | Ho is True) = α = alpha = **Significance Level**

A **TYPE II Error** occurs when we fail to Reject Ho when, in fact, Ho is False. In this case **we fail to reject a false null hypothesis.**

- P(TYPE II Error) = P(Fail to Reject Ho | Ho is False) = β = beta

When our significance level is 5%, we are saying that we will allow ourselves to make a Type I error less than 5% of the time. In the long run, if we repeat the process, 5% of the time we will find a p-value < 0.05 when in fact the null hypothesis was true.

In this case, our data represent a rare occurrence which is unlikely to happen but is still possible. For example, suppose we toss a coin 10 times and obtain 10 heads, this is unlikely for a fair coin but not impossible. We might conclude the coin is unfair when in fact we simply saw a very rare event for this fair coin.

Our testing procedure CONTROLS for the Type I error when we set a pre-determined value for the significance level.

Notice that these probabilities are conditional probabilities. This is one more reason why conditional probability is an important concept in statistics.

Unfortunately, calculating the probability of a Type II error requires us to know the truth about the population. In practice we can only calculate this probability using a series of "what if" calculations which depend upon the type of problem.

Caution

Comment: As you initially read through the examples below, focus on the broad concepts instead of the small details. It is not important to understand how to calculate these values yourself at this point.

- Try to understand the pictures we present. Which pictures represent an assumed null hypothesis and which represent an alternative?
- It may be useful to come back to this page (and the activities here) after you have reviewed the rest of the section on hypothesis testing and have worked a few problems yourself.

Interactive Applet: Statistical Significance

Here are two examples of using an older version of this applet. It looks slightly different but the same settings and options are available in the version above.

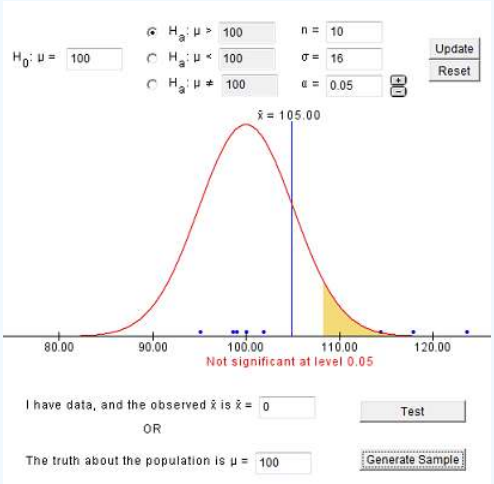
In both cases we will consider IQ scores.

Our null hypothesis is that the true mean is 100. Assume the standard deviation is 16 and we will specify a significance level of 5%.

EXAMPLE:

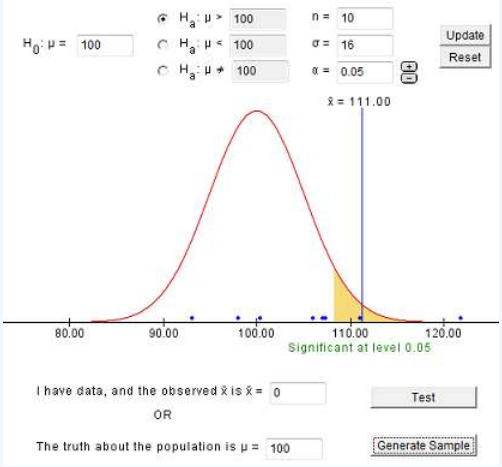
In this example we will specify that the true mean is indeed 100 so that the null hypothesis is true. Most of the time (95%), when we generate a sample, we should fail to reject the null hypothesis since the null hypothesis is indeed true.

Here is one sample that results in a correct decision:



In the sample above, we obtain an x-bar of 105, which is drawn on the distribution which assumes μ (mu) = 100 (the null hypothesis is true). Notice the sample is shown as blue dots along the x-axis and the shaded region shows for which values of x-bar we would reject the null hypothesis. In other words, we would reject Ho whenever the x-bar falls in the shaded region.

Enter the same values and generate samples until you obtain a Type I error (you falsely reject the null hypothesis). You should see something like this:



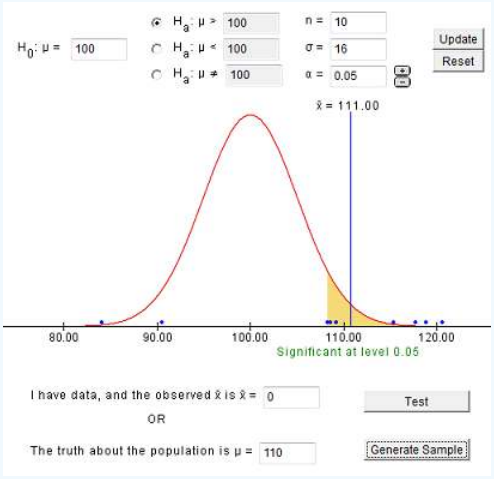
If you were to generate 100 samples, you should have around 5% where you rejected Ho. These would be samples which would result in a Type I error.

The previous example illustrates a correct decision and a Type I error when the null hypothesis is true. The next example illustrates a correct decision and Type II error when the null hypothesis is false. In this case, we must specify the true population mean.

✔ EXAMPLE:

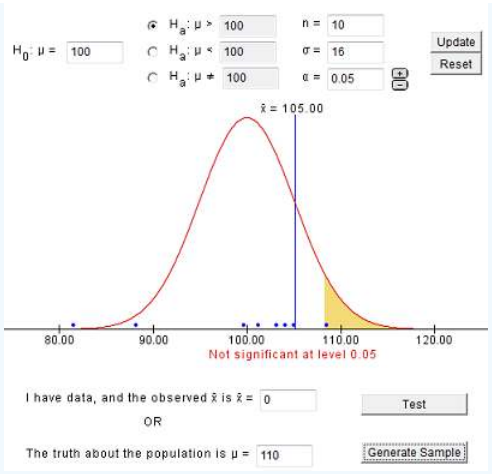
Let's suppose we are sampling from an honors program and that the true mean IQ for this population is 110. We do not know the probability of a Type II error without more detailed calculations.

Let's start with a sample which results in a correct decision.



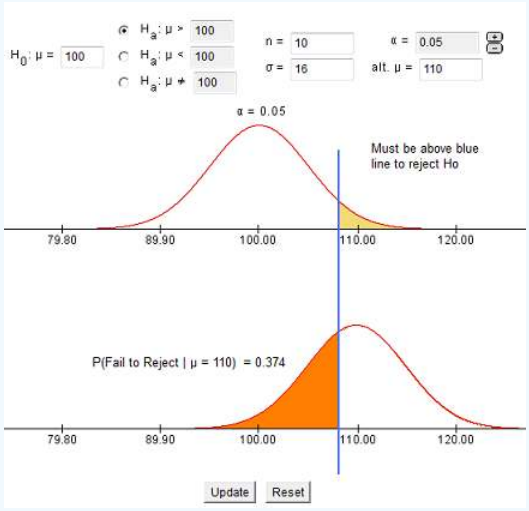
In the sample above, we obtain an x-bar of 111, which is drawn on the distribution which assumes μ (mu) = 100 (the null hypothesis is true).

Enter the same values and generate samples until you obtain a Type II error (you fail to reject the null hypothesis). You should see something like this:



You should notice that in this case (when Ho is false), it is easier to obtain an incorrect decision (a Type II error) than it was in the case where Ho is true. If you generate 100 samples, you can approximate the probability of a Type II error.

We can find the probability of a Type II error by visualizing both the assumed distribution and the true distribution together. The image below is adapted from an applet we will use when we discuss the power of a statistical test.

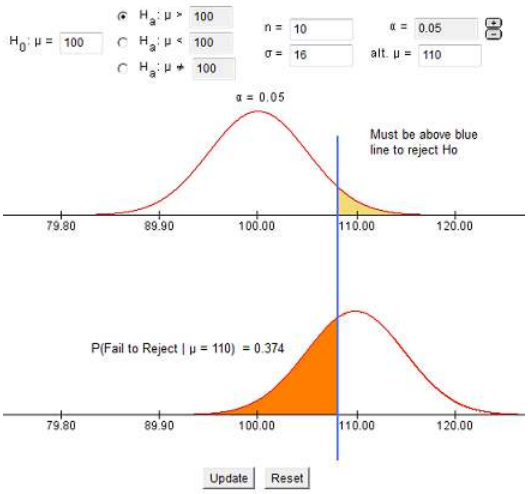


There is a 37.4% chance that, in the long run, we will make a Type II error and fail to reject the null hypothesis when in fact the true mean IQ is 110 in the population from which we sample our 10 individuals.

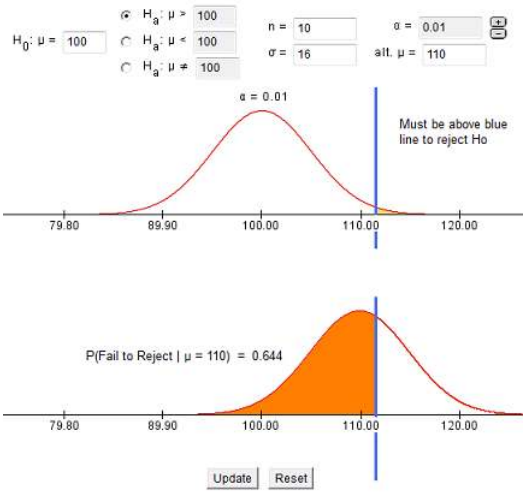
Can you visualize what will happen if the true population mean is really 115 or 108? When will the Type II error increase? When will it decrease? We will look at this idea again when we discuss the concept of power in hypothesis tests.

Comments:

- It is important to note that there is a trade-off between the probability of a Type I and a Type II error. If we decrease the probability of one of these errors, the probability of the other will increase! The practical result of this is that if we require stronger evidence to reject the null hypothesis (smaller significance level = probability of a Type I error), we will increase the chance that we will be unable to reject the null hypothesis when in fact Ho is false (increases the probability of a Type II error).
- When α (alpha) = 0.05 we obtained a Type II error probability of 0.374 = β = beta



- When α (alpha) = 0.01 (smaller than before) we obtain a Type II error probability of 0.644 = β = beta (larger than before)



- As the blue line in the picture moves farther right, the significance level (α, alpha) is decreasing and the Type II error probability is increasing.
- As the blue line in the picture moves farther left, the significance level (α, alpha) is increasing and the Type II error probability is decreasing

Let's return to our very first example and define these two errors in context.

EXAMPLE:

A case of suspected cheating on an exam is brought in front of the disciplinary committee at a certain university.

There are **two** opposing **claims** in this case:

- Ho = The **student's claim:** I did not cheat on the exam.
- Ha = The **instructor's claim:** The student did cheat on the exam.

Adhering to the principle **"innocent until proven guilty,"** the committee asks the instructor for **evidence** to support his claim.

There are four possible outcomes of this process. There are two possible correct decisions:

- The student did cheat on the exam and the instructor brings enough evidence to reject Ho and conclude the student did cheat on the exam. This is a CORRECT decision!

- The student did not cheat on the exam and the instructor fails to provide enough evidence that the student did cheat on the exam. This is a CORRECT decision!

Both the correct decisions and the possible errors are fairly easy to understand but with the errors, you must be careful to identify and define the two types correctly.

TYPE I Error: Reject Ho when Ho is True

- The student did not cheat on the exam but the instructor brings enough evidence to reject Ho and conclude the student cheated on the exam. This is a Type I Error.

TYPE II Error: Fail to Reject Ho when Ho is False

- The student did cheat on the exam but the instructor fails to provide enough evidence that the student cheated on the exam. This is a Type II Error.

In most situations, including this one, it is more "acceptable" to have a Type II error than a Type I error. Although allowing a student who cheats to go unpunished might be considered a very bad problem, punishing a student for something he or she did not do is usually considered to be a more severe error. This is one reason we control for our Type I error in the process of hypothesis testing.

Did I Get This?: Type I and Type II Errors (in context)

Comment:

- The probabilities of Type I and Type II errors are closely related to the concepts of sensitivity and specificity that we discussed previously. Consider the following hypotheses:

Ho: The individual does not have diabetes (status quo, nothing special happening)

Ha: The individual does have diabetes (something is going on here)

In this setting:

When someone tests positive for diabetes we would reject the null hypothesis and conclude the person has diabetes (we may or may not be correct!).

When someone tests negative for diabetes we would fail to reject the null hypothesis so that we fail to conclude the person has diabetes (we may or may not be correct!)

Let's take it one step further:

Sensitivity = P(Test + | Have Disease) which in this setting equals
P(Reject Ho | Ho is False) = 1 – P(Fail to Reject Ho | Ho is False) = 1 – β = 1 – beta

Specificity = P(Test – | No Disease) which in this setting equals
P(Fail to Reject Ho | Ho is True) = 1 – P(Reject Ho | Ho is True) = 1 – α = 1 – alpha

Notice that sensitivity and specificity relate to the probability of making a correct decision whereas α (alpha) and β (beta) relate to the probability of making an incorrect decision.

Usually α (alpha) = 0.05 so that the specificity listed above is 0.95 or 95%.

Next, we will see that the sensitivity listed above is the **power** of the hypothesis test!

Reasons for a Type I Error in Practice

Assuming that you have obtained a quality sample:

- The reason for a Type I error is random chance.
- When a Type I error occurs, our observed data represented a rare event which indicated evidence in favor of the alternative hypothesis even though the null hypothesis was actually true.

Reasons for a Type II Error in Practice

Again, assuming that you have obtained a quality sample, now we have a few possibilities depending upon the true difference that exists.

- The sample size is too small to detect an important difference. This is the worst case, you should have obtained a larger sample. In this situation, you may notice that the effect seen in the sample seems PRACTICALLY significant and yet the p-value is not small enough to reject the null hypothesis.
- The sample size is reasonable for the important difference but the true difference (which might be somewhat meaningful or interesting) is smaller than your test was capable of detecting. This is tolerable as you were not interested in being able to detect this difference when you began your study. In this situation, you may notice that the effect seen in the sample seems to have some potential for practical significance.
- The sample size is more than adequate, the difference that was not detected is meaningless in practice. This is not a problem at all and is in effect a "correct decision" since the difference you did not detect would have no practical meaning.
- Note: We will discuss the idea of practical significance later in more detail.

Power of a Hypothesis Test

It is often the case that we truly wish to prove the alternative hypothesis. It is reasonable that we would be interested in the probability of correctly rejecting the null hypothesis. In other words, the probability of rejecting the null hypothesis, when in fact the null hypothesis is false. This can also be thought of as the probability of being able to detect a (pre-specified) difference of interest to the researcher.

Let's begin with a realistic example of how power can be described in a study.

✔ EXAMPLE:

In a clinical trial to study two medications for weight loss, we have an 80% chance to detect a difference in the weight loss between the two medications of 10 pounds. In other words, the power of the hypothesis test we will conduct is 80%.

In other words, if one medication comes from a population with an average weight loss of 25 pounds and the other comes from a population with an average weight loss of 15 pounds, we will have an 80% chance to detect that difference using the sample we have in our trial.

If we were to repeat this trial many times, 80% of the time we will be able to reject the null hypothesis (that there is no difference between the medications) and 20% of the time we will fail to reject the null hypothesis (and make a Type II error!).

The difference of 10 pounds in the previous example, is often called the **effect size**. The measure of the effect differs depending on the particular test you are conducting but is always some measure related to the true effect in the population. In this example, it is the difference between two population means.

Recall the definition of a Type II error:

A **TYPE II Error** occurs when we fail to Reject Ho when, in fact, Ho is False. In this case **we fail to reject a false null hypothesis.**

P(TYPE II Error) = P(Fail to Reject Ho | Ho is False) = β = beta

Notice that P(Reject Ho | Ho is False) = 1 – P(Fail to Reject Ho | Ho is False) = 1 – β = 1- beta.

The **POWER** of a hypothesis test is the **probability of rejecting the null hypothesis when the null hypothesis is false**. This can also be stated as the **probability of correctly rejecting the null hypothesis**.

POWER = P(Reject Ho | Ho is False) = 1 – β = 1 – beta

Power is the test's ability to correctly reject the null hypothesis. **A test with high power has a good chance of being able to detect the difference of interest to us, if it exists**.

As we mentioned on the bottom of the previous page, this can be thought of as the sensitivity of the hypothesis test if you imagine Ho = No disease and Ha = Disease.

Factors Affecting the Power of a Hypothesis Test

The power of a hypothesis test is affected by numerous quantities (similar to the margin of error in a confidence interval).

Assume that the null hypothesis is false for a given hypothesis test. All else being equal, we have the following:

- Larger samples result in a greater chance to reject the null hypothesis which means an increase in the power of the hypothesis test.
- If the **effect size** is larger, it will become easier for us to detect. This results in a greater chance to reject the null hypothesis which means an increase in the power of the hypothesis test. The effect size varies for each test and is usually closely related to the difference between the hypothesized value and the true value of the parameter under study.
- From the relationship between the probability of a Type I and a Type II error (as α (alpha) decreases, β (beta) increases), we can see that as α (alpha) decreases, Power = 1 – β = 1 – beta also decreases.
- There are other mathematical ways to change the power of a hypothesis test, such as changing the population standard deviation; however, these are not quantities that we can usually control so we will not discuss them here.

Caution

In practice, we specify a significance level and a desired power to detect a difference which will have practical meaning to us and this determines the sample size required for the experiment or study.

For most grants involving statistical analysis, power calculations must be completed to illustrate that the study will have a reasonable chance to detect an important effect. Otherwise, the money spent on the study could be wasted. The goal is usually to have a power close to 80%.

For example, if there is only a 5% chance to detect an important difference between two treatments in a clinical trial, this would result in a waste of time, effort, and money on the study since, when the alternative hypothesis is true, the chance a treatment effect can be found is very small.

Comment:

- In order to calculate the power of a hypothesis test, we must specify the "truth." As we mentioned previously when discussing Type II errors, in practice we can only calculate this probability using a series of "what if" calculations which depend upon the type of problem.

The following activity involves working with an interactive applet to study power more carefully.

Learn by Doing: Power of Hypothesis Tests

The following reading is an excellent discussion about Type I and Type II errors.

(Optional) Outside Reading: A Good Discussion of Power (≈ 2500 words)

We will not be asking you to perform power calculations manually. You may be asked to use online calculators and applets. Most statistical software packages offer some ability to complete power calculations. There are also many online calculators for power and sample size on the internet, for example, Russ Lenth's power and sample-size page.

Proportions (Introduction & Step 1)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.33: In a given context, distinguish between situations involving a population proportion and a population mean and specify the correct null and alternative hypothesis for the scenario.

Learning Objectives

LO 4.34: Carry out a complete hypothesis test for a population proportion by hand.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Video

Video: Proportions (Introduction & Step 1) (7:18)

Now that we understand the process of hypothesis testing and the logic behind it, we are ready to start learning about specific statistical tests (also known as significance tests).

The first test we are going to learn is the test about the population proportion (p).

This test is widely known as the **"z-test for the population proportion (p)."**

Introduction

We will understand later where the "z-test" part is coming from.

This will be the only type of problem you will complete entirely "by-hand" in this course. Our goal is to use this example to give you the tools you need to understand how this process works. After working a few problems, you should review the earlier material again. You will likely need to review the terminology and concepts a few times before you fully understand the process.

In reality, you will often be conducting more complex statistical tests and allowing software to provide the p-value. In these settings it will be important to know what test to apply for a given situation and to be able to explain the results in context.

Review: Types of Variables

When we conduct a test about a population proportion, we are working with a categorical variable. Later in the course, after we have learned a variety of hypothesis tests, we will need to be able to identify which test is appropriate for which situation. Identifying the variable as categorical or quantitative is an important component of choosing an appropriate hypothesis test.

Learn by Doing: Review Types of Variables

One Sample Z-Test for a Population Proportion

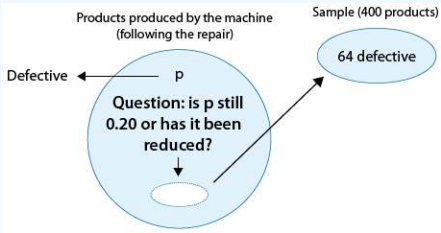
In this part of our discussion on hypothesis testing, we will go into details that we did not go into before. More specifically, we will use this test to introduce the idea of a **test statistic**, and details about **how p-values are calculated**.

Let's start by introducing the three examples, which will be the leading examples in our discussion. Each example is followed by a figure illustrating the information provided, as well as the question of interest.

EXAMPLE:

A machine is known to produce 20% defective products, and is therefore sent for repair. After the machine is repaired, 400 products produced by the machine are chosen at random and 64 of them are found to be defective. Do the data provide enough evidence that the proportion of defective products produced by the machine (p) has been **reduced** as a result of the repair?

The following figure displays the information, as well as the question of interest:



The question of interest helps us formulate the null and alternative hypotheses in terms of p, the proportion of defective products produced by the machine following the repair:

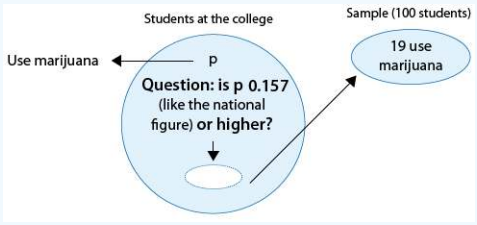
Ho: p = 0.20 (No change; the repair did not help).

Ha: p < 0.20 (The repair was effective at reducing the proportion of defective parts).

EXAMPLE:

There are rumors that students at a certain liberal arts college are more inclined to use drugs than U.S. college students in general. Suppose that in a simple random sample of 100 students from the college, 19 admitted to marijuana use. Do the data provide enough evidence to conclude that the proportion of marijuana users among the students in the college (p) is **higher** than the national proportion, which is 0.157? (This number is reported by the Harvard School of Public Health.)

Again, the following figure displays the information as well as the question of interest:



As before, we can formulate the null and alternative hypotheses in terms of p, the proportion of students in the college who use marijuana:

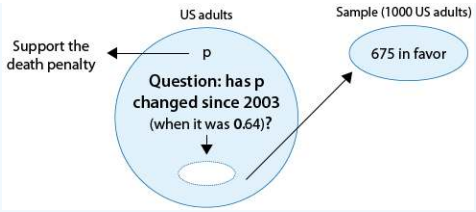
Ho: p = 0.157 (same as among all college students in the country).

Ha: p > 0.157 (higher than the national figure).

EXAMPLE:

Polls on certain topics are conducted routinely in order to monitor changes in the public's opinions over time. One such topic is the death penalty. In 2003 a poll estimated that 64% of U.S. adults support the death penalty for a person convicted of murder. In a more recent poll, 675 out of 1,000 U.S. adults chosen at random were in favor of the death penalty for convicted murderers. Do the results of this poll provide evidence that the proportion of U.S. adults who support the death penalty for convicted murderers (p) **changed** between 2003 and the later poll?

Here is a figure that displays the information, as well as the question of interest:



Again, we can formulate the null and alternative hypotheses in term of p, the proportion of U.S. adults who support the death penalty for convicted murderers.

Ho: p = 0.64 (No change from 2003).

Ha: p ≠ 0.64 (Some change since 2003).

Learn by Doing: Proportions (Overview)

Did I Get This?: Proportions (Overview)

Recall that there are basically 4 steps in the process of hypothesis testing:

- **STEP 1:** State the appropriate null and alternative hypotheses, Ho and Ha.
- **STEP 2:** Obtain a random sample, collect relevant data, and **check whether the data meet the conditions under which the test can be used**. If the conditions are met, summarize the data using a test statistic.
- **STEP 3:** Find the p-value of the test.
- **STEP 4:** Based on the p-value, decide whether or not the results are statistically significant and **draw your conclusions in context.**
- **Note:** In practice, we should always consider the practical significance of the results as well as the statistical significance.

We are now going to go through these steps as they apply to the hypothesis testing for the population proportion p. It should be noted that even though the details will be specific to this particular test, some of the ideas that we will add apply to hypothesis testing in general.

Step 1. Stating the Hypotheses

Here again are the three set of hypotheses that are being tested in each of our three examples:

EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?

- **Ho:** p = 0.20 (No change; the repair did not help).
- **Ha:** p < 0.20 (The repair was effective at reducing the proportion of defective parts).

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?

- **Ho:** p = 0.157 (same as among all college students in the country).
- **Ha:** p > 0.157 (higher than the national figure).

EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?

- **Ho:** p = 0.64 (No change from 2003).
- **Ha:** p ≠ 0.64 (Some change since 2003).

The null hypothesis always takes the form:

- Ho: p = some value

and the alternative hypothesis takes one of the following three forms:

- Ha: p < that value (like in example 1) **or**
- Ha: p > that value (like in example 2) **or**
- Ha: p ≠ that value (like in example 3).

Note that it was quite clear from the context which form of the alternative hypothesis would be appropriate. The value that is specified in the null hypothesis is called the **null value**, and is generally denoted by p_0. We can say, therefore, that in general the null hypothesis about the population proportion (p) would take the form:

- Ho: $p = p_0$

We write Ho: $p = p_0$ to say that we are making the hypothesis that the population proportion has the value of p_0. In other words, p is the unknown population proportion and p_0 is the number we think p might be for the given situation.

The alternative hypothesis takes one of the following three forms (depending on the context):

- Ha: $p < p_0$ **(one-sided)**
- Ha: $p > p_0$ **(one-sided)**
- Ha: $p \neq p_0$ **(two-sided)**

The first two possible forms of the alternatives (where the = sign in Ho is challenged by < or >) are called **one-sided alternatives**, and the third form of alternative (where the = sign in Ho is challenged by ≠) is called a **two-sided alternative.** To understand the intuition behind these names let's go back to our examples.

Example 3 (death penalty) is a case where we have a two-sided alternative:

- **Ho:** p = 0.64 (No change from 2003).
- **Ha:** p ≠ 0.64 (Some change since 2003).

In this case, in order to reject Ho and accept Ha we will need to get a sample proportion of death penalty supporters which is very different from 0.64 **in either direction,** either much larger or much smaller than 0.64.

In example 2 (marijuana use) we have a one-sided alternative:

- **Ho:** p = 0.157 (same as among all college students in the country).
- **Ha:** p > 0.157 (higher than the national figure).

Here, in order to reject Ho and accept Ha we will need to get a sample proportion of marijuana users which is much **higher** than 0.157.

Similarly, in example 1 (defective products), where we are testing:

- **Ho:** p = 0.20 (No change; the repair did not help).
- **Ha:** p < 0.20 (The repair was effective at reducing the proportion of defective parts).

in order to reject Ho and accept Ha, we will need to get a sample proportion of defective products which is much **smaller** than 0.20.

Learn by Doing: State Hypotheses (Proportions)

Did I Get This?: State Hypotheses (Proportions)

Proportions (Step 2)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.33: In a given context, distinguish between situations involving a population proportion and a population mean and specify the correct null and alternative hypothesis for the scenario.

Learning Objectives

LO 4.34: Carry out a complete hypothesis test for a population proportion by hand.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Video

Video: Proportions (Step 2) (12:38)

Step 2. Collect Data, Check Conditions, and Summarize Data

After the hypotheses have been stated, the next step is to obtain a **sample** (on which the inference will be based), **collect relevant data**, and **summarize** them.

It is extremely important that our sample is representative of the population about which we want to draw conclusions. This is ensured when the sample is chosen at **random.** Beyond the practical issue of ensuring representativeness, choosing a random sample has theoretical importance that we will mention later.

In the case of hypothesis testing for the population proportion (p), we will collect data on the relevant categorical variable from the individuals in the sample and start by calculating the sample proportion p-hat (the natural quantity to calculate when the parameter of interest is p).

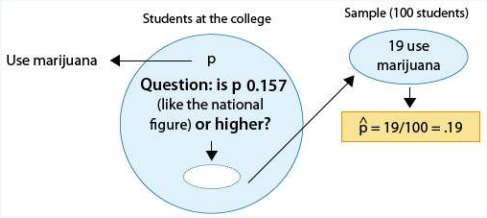
Let's go back to our three examples and add this step to our figures.

EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?



EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?

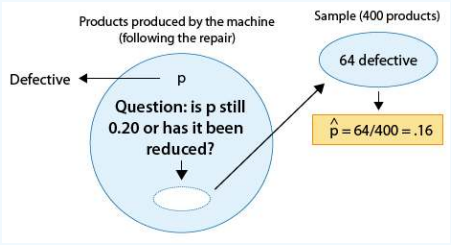
As we mentioned earlier without going into details, when we summarize the data in hypothesis testing, we go a step beyond calculating the sample statistic and summarize the data with a **test statistic**. Every test has a test statistic, which to some degree captures the essence of the test. In fact, the p-value, which so far we have looked upon as "the king" (in the sense that everything is determined by it), is actually determined by (or derived from) the test statistic. We will now introduce the test statistic.

The test statistic is a measure of how far the sample proportion p-hat is from the null value p_0, the value that the null hypothesis claims is the value of p. In other words, since p-hat is what the data estimates p to be, the test statistic can be viewed as a measure of the "distance" between what the data tells us about p and what the null hypothesis claims p to be.

Let's use our examples to understand this:

EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?



The parameter of interest is p, the proportion of defective products following the repair.

The data estimate p to be p-hat = 0.16

The null hypothesis claims that p = 0.20

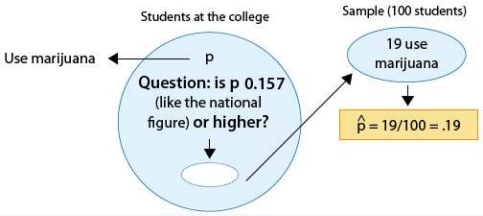
The data are therefore 0.04 (or 4 percentage points) below the null hypothesis value.

It is hard to evaluate whether this difference of 4% in defective products is enough evidence to say that the repair was effective at reducing the proportion of defective products, but clearly, the larger the difference, the more evidence it is against the null hypothesis. So if, for example, our sample proportion of defective products had been, say, 0.10 instead of 0.16, then I think you

would all agree that cutting the proportion of defective products in half (from 20% to 10%) would be extremely strong evidence that the repair was effective at reducing the proportion of defective products.

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?



The parameter of interest is p, the proportion of students in a college who use marijuana.

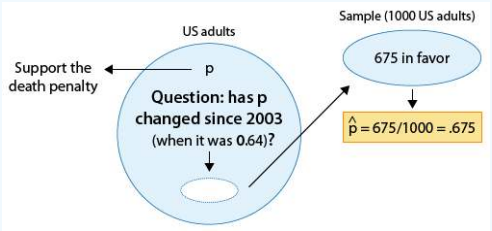
The data estimate p to be p-hat = 0.19

The null hypothesis claims that p = 0.157

The data are therefore 0.033 (or 3.3. percentage points) above the null hypothesis value.

EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?



The parameter of interest is p, the proportion of U.S. adults who support the death penalty for convicted murderers.

The data estimate p to be p-hat = 0.675

The null hypothesis claims that p = 0.64

There is a difference of 0.035 (or 3.5. percentage points) between the data and the null hypothesis value.

The problem with looking only at the difference between the sample proportion, p-hat, and the null value, p_0 is that we have not taken into account the variability of our estimator p-hat which, as we know from our study of sampling distributions, depends on the sample size.

For this reason, the test statistic cannot simply be the difference between p-hat and p_0, but must be some form of that formula that accounts for the sample size. In other words, we need to somehow standardize the difference so that comparison between different situations will be possible. We are very close to revealing the test statistic, but before we construct it, let's be reminded of the following two facts from probability:

Fact 1: When we take a random sample of size n from a population with population proportion p, then

$$\hat{p} \text{ is normally distributed with a mean of } \mu_{\hat{p}} = p$$

$$\text{and a standard deviation } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\text{as long as } np \geq 10 \text{ and } n(1-p) \geq 10$$

Fact 2: The z-score of any normal value (a value that comes from a normal distribution) is calculated by finding the difference between the value and the mean and then dividing that difference by the standard deviation (of the normal distribution associated with the value). The z-score represents how many standard deviations below or above the mean the value is.

Thus, our test statistic should be **a measure** of how far the sample proportion p-hat is from the null value p_0 **relative** to the variation of p-hat (as measured by the standard error of p-hat).

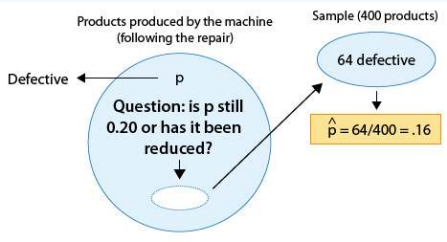
Recall that the **standard error** is the **standard deviation of the sampling distribution** for a given statistic. For p-hat, we know the following:

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal if np ≥ 10 and n(1 - p) ≥ 10

To find the p-value, we will need to determine how surprising our value is assuming the null hypothesis is true. We already have the tools needed for this process from our study of sampling distributions as represented in the table above.

✔ EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?

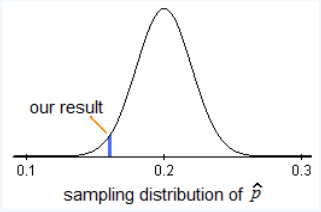


If we assume the null hypothesis is true, we can specify that the center of the distribution of all possible values of p-hat from samples of size 400 would be 0.20 (our null value).

We can calculate the standard error, assuming p = 0.20 as

$$\sqrt{\frac{p_0\,(1-p_0)}{n}} = \sqrt{\frac{0.2(1-0.2)}{400}} = 0.02$$

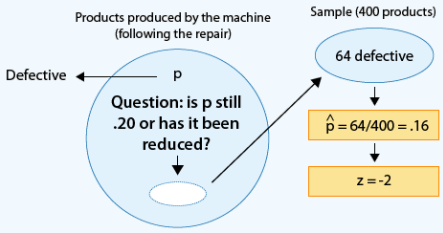
The following picture represents the sampling distribution of all possible values of p-hat of samples of size 400, assuming the true proportion p is 0.20 and our other requirements for the sampling distribution to be normal are met (we will review these during the next step).



In order to calculate probabilities for the picture above, we would need to find the z-score associated with our result.

This z-score is the **test statistic**! In this example, the numerator of our z-score is the difference between p-hat (0.16) and null value (0.20) which we found earlier to be -0.04. The denominator of our z-score is the standard error calculated above (0.02) and thus quickly we find the z-score, our test statistic, to be -2.

The sample proportion based upon this data is 2 standard errors below the null value.



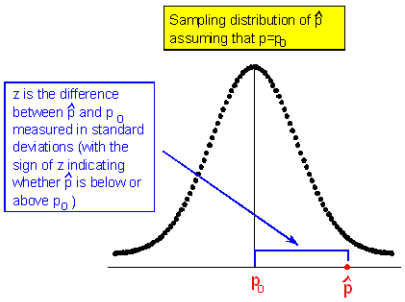
Hopefully you now understand more about the reasons we need probability in statistics!!

Now we will formalize the definition and look at our remaining examples before moving on to the next step, which will be to determine if a normal distribution applies and calculate the p-value.

Test Statistic for Hypothesis Tests for One Proportion is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

It represents the difference between the sample proportion and the null value, measured in standard deviations (standard error of p-hat).



The picture above is a representation of the sampling distribution of p-hat assuming p = p_0. In other words, this is a model of how p-hat behaves if we are drawing random samples from a population for which Ho is true.

Notice the center of the sampling distribution is at p_0, which is the hypothesized proportion given in the null hypothesis (Ho: p = p_0.) We could also mark the axis in standard error units,

$$\sqrt{\frac{p_0\,(1-p_0)}{n}}$$

For example, if our null hypothesis claims that the proportion of U.S. adults supporting the death penalty is 0.64, then the sampling distribution is drawn as if the null is true. We draw a normal distribution centered at 0.64 (p_0) with a standard error dependent on sample size,

$$\sqrt{\frac{0.64(1-0.64)}{n}}.$$

Important Comment:

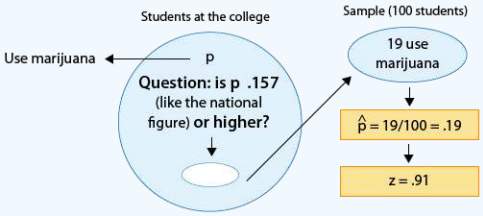
- Note that under the assumption that Ho is true (and if the conditions for the sampling distribution to be normal are satisfied) the test statistic follows a N(0,1) (standard normal) distribution. Another way to say the same thing which is quite common is: “The null distribution of the test statistic is N(0,1).”

By “null distribution,” we mean the distribution under the assumption that Ho is true. As we’ll see and stress again later, the null distribution of the test statistic is what the calculation of the p-value is based on.

Let’s go back to our remaining two examples and find the test statistic in each case:

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?



Since the null hypothesis is Ho: p = 0.157, the standardized (z) score of p-hat = 0.19 is

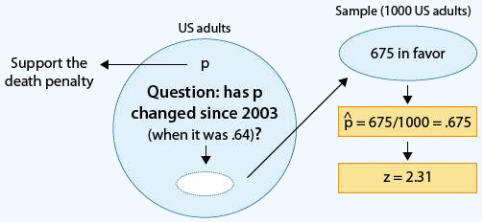
$$z = \frac{0.19 - 0.157}{\sqrt{\frac{0.157(1-0.157)}{100}}} \approx 0.91$$

This is the value of the test statistic for this example.

We interpret this to mean that, assuming that Ho is true, the sample proportion p-hat = 0.19 is 0.91 standard errors above the null value (0.157).

EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?



Since the null hypothesis is Ho: p = 0.64, the standardized (z) score of p-hat = 0.675 is

$$z = \frac{0.675 - 0.64}{\sqrt{\frac{0.64(1-0.64)}{1000}}} \approx 2.31$$

This is the value of the test statistic for this example.

We interpret this to mean that, assuming that Ho is true, the sample proportion p-hat = 0.675 is 2.31 standard errors above the null value (0.64).

Learn by Doing: Proportions (Step 2)

Comments about the Test Statistic:

- We mentioned earlier that to some degree, the test statistic captures the essence of the test. In this case, the test statistic measures the difference between p-hat and p_0 in standard errors. This is exactly what this test is about. Get data, and look at the discrepancy between what the data estimates p to be (represented by p-hat) and what Ho claims about p (represented by p_0).
- You can think about this test statistic as a measure of evidence in the data against Ho. The larger the test statistic, the "further the data are from Ho" and therefore the more evidence the data provide against Ho.

Learn by Doing: Proportions (Step 2) Understanding the Test Statistic

Did I Get This?: Proportions (Step 2)

Comments:

- It should now be clear why this test is commonly known as **the z-test for the population proportion**. The name comes from the fact that it is based on a test statistic that is a *z-score.*
- Recall fact 1 that we used for constructing the z-test statistic. Here is part of it again:

When we take a **random** sample of size n from a population with population proportion p_0, the possible values of the sample proportion p-hat (**when certain conditions are met**) have approximately a normal distribution with a mean of p_0… and a standard deviation of

$$\sqrt{\frac{p_0(1-p_0)}{n}}$$

This result provides the theoretical justification for constructing the test statistic the way we did, and therefore the assumptions under which this result holds (in bold, above) are the conditions that our data need to satisfy so that we can use this test. These two conditions are:

i. The sample has to be random.

ii. The conditions under which the sampling distribution of p-hat is normal are met. In other words:

$$np_0 \geq 10$$

$$n(1-p_0) \geq 10$$

- Here we will pause to say more about condition (i.) above, the need for a random sample. In the Probability Unit we discussed sampling plans based on probability (such as a simple random sample, cluster, or stratified sampling) that produce a non-biased sample, which can be safely used in order to make inferences about a population. We noted in the Probability Unit that, in practice, other (non-random) sampling techniques are sometimes used when random sampling is not feasible. It is important though, when these techniques are used, to be aware of the type of bias that they introduce, and thus the limitations of the conclusions that can be drawn from them. For our purpose here, we will focus on one such practice, the situation in which a sample is not really chosen randomly, but in the context of the categorical variable that is being studied, the sample is regarded as random. For example, say that you are interested in the proportion of students at a certain college who suffer from seasonal allergies. For that purpose, the students in a large engineering class could be considered as a random sample, since there is nothing about being in an engineering class that makes you more or less likely to suffer from seasonal allergies. Technically, the engineering class is a convenience sample, but it is treated as a random sample in the context of this categorical variable. On the other hand, if you are interested in the proportion of students in the college who have math anxiety, then the class of engineering students clearly could not possibly be viewed as a random sample, since engineering students probably have a much lower incidence of math anxiety than the college population overall.

Learn by Doing: Proportions (Step 2) Valid or Invalid Sampling?

Let's check the conditions in our three examples.

✔ EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?

i. The 400 products were chosen at random.

ii. n = 400, p_0 = 0.2 and therefore:

$$np_0 = 400(0.2) = 80 \geq 10$$

$$n\,(1-p_0) = 400(1-0.2) = 320 \geq 10$$

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?

i. The 100 students were chosen at random.

ii. n = 100, p_0 = 0.157 and therefore:

$$np_0 = 100(0.157) = 15.7 \geq 10$$
$$n(1-p_0) = 100(1-0.157) = 84.3 \geq 10$$

EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?

i. The 1000 adults were chosen at random.

ii. n = 1000, p_0 = 0.64 and therefore:

$$np_0 = 1000(0.64) = 640 \geq 10$$
$$n(1-p_0) = 1000(1-0.64) = 360 \geq 10$$

Learn by Doing: Proportions (Step 2) Verify Conditions

Checking that our data satisfy the conditions under which the test can be reliably used is a very important part of the hypothesis testing process. Be sure to consider this for every hypothesis test you conduct in this course and certainly in practice.

The Four Steps in Hypothesis Testing

- **STEP 1:** State the appropriate null and alternative hypotheses, Ho and Ha.
- **STEP 2:** Obtain a random sample, collect relevant data, and **check whether the data meet the conditions under which the test can be used**. If the conditions are met, summarize the data using a test statistic.
- **STEP 3:** Find the p-value of the test.
- **STEP 4:** Based on the p-value, decide whether or not the results are statistically significant and **draw your conclusions in context.**
- **Note:** In practice, we should always consider the practical significance of the results as well as the statistical significance.

With respect to the z-test, the population proportion that we are currently discussing we have:

Step 1: Completed

Step 2: Completed

Step 3: This is what we will work on next.

Proportions (Step 3)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.33: In a given context, distinguish between situations involving a population proportion and a population mean and specify the correct null and alternative hypothesis for the scenario.

Learning Objectives

LO 4.34: Carry out a complete hypothesis test for a population proportion by hand.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Learning Objectives

LO 6.27: Explain what the p-value is and how it is used to draw conclusions.

Video

Video: Proportions (Step 3) (14:46)

Calculators and Tables

Step 3. Finding the P-value of the Test

So far we've talked about the p-value at the intuitive level: understanding what it is (or what it measures) and how we use it to draw conclusions about the statistical significance of our results. We will now go more deeply into how the p-value is calculated.

It should be mentioned that eventually we will rely on technology to calculate the p-value for us (as well as the test statistic), but in order to make intelligent use of the output, it is important to first **understand** the details, and only then let the computer do the calculations for us. Again, our goal is to use this simple example to give you the tools you need to understand the process entirely. Let's start.

Recall that so far we have said that the p-value is the probability of obtaining data like those observed assuming that Ho is true. Like the test statistic, the p-value is, therefore, a measure of the evidence against Ho. In the case of the **test statistic,** the **larger** it is in magnitude (positive or negative), the further p-hat is from p_0, the **more evidence we have against Ho.** In the case of the **p-value**, it is the opposite; the **smaller** it is, the more unlikely it is to get data like those observed when Ho is true, the **more evidence it is against Ho**. One can actually draw conclusions in hypothesis testing just using the test statistic, and as we'll see the p-value is, in a sense, just another way of looking at the test statistic. The reason that we actually take the extra step in this course and derive the p-value from the test statistic is that even though in this case (the test about the population proportion) and some other tests, the value of the test statistic has a very clear and intuitive interpretation, there are some tests where its value is not as easy to interpret. On the other hand, the p-value keeps its intuitive appeal across **all** statistical tests.

How is the p-value calculated?

Intuitively, the p-value is the **probability** of observing **data like those observed** assuming that Ho is true. Let's be a bit more formal:

- Since this is a probability question about the **data**, it makes sense that the calculation will involve the data summary, the **test statistic.**
- What do we mean by **"like"** those observed? By "like" we mean **"as extreme or even more extreme."**

Putting it all together, we get that in **general:**

The **p-value** is the **probability of observing a test statistic as extreme as that observed (or even more extreme) assuming that the null hypothesis is true.**

By **"extreme"** we mean extreme **in the direction(s) of the alternative** hypothesis.

Specifically, for the z-test for the population proportion:

1. If the alternative hypothesis is Ha: $p < p_0$ **(less than)**, then "extreme" means **small or less than**, and the p-value is: The probability of observing a test statistic **as small as that observed or smaller** if the null hypothesis is true.
2. If the alternative hypothesis is Ha: $p > p_0$ **(greater than)**, then "extreme" means **large or greater than**, and the p-value is: The probability of observing a test statistic **as large as that observed or larger** if the null hypothesis is true.
3. If the alternative is Ha: $p \neq p_0$ **(different from)**, then "extreme" means extreme in either direction **either small or large (i.e., large in magnitude) or just different from**, and the p-value therefore is: The probability of observing a test statistic **as large in magnitude as that observed or larger** if the null hypothesis is true.(Examples: If z = -2.5: p-value = probability of observing a test statistic as small as -2.5 or smaller or as large as 2.5 or larger. If z = 1.5: p-value = probability of observing a test statistic as large as 1.5 or larger, or as small as -1.5 or smaller.)

OK, hopefully that makes (some) sense. But how do we actually calculate it?

Recall the important comment from our discussion about our test statistic,

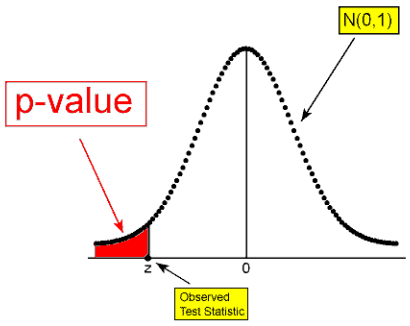
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

which said that when the null hypothesis is true (i.e., when $p = p_0$), the possible values of our test statistic follow a standard normal (N(0,1), denoted by Z) distribution. Therefore, the p-value calculations (which assume that Ho is true) are simply standard normal distribution calculations for the 3 possible alternative hypotheses.

Alternative Hypothesis is "Less Than"

The probability of observing a test statistic as **small as that observed or smaller**, assuming that the values of the test statistic follow a standard normal distribution. We will now represent this probability in symbols and also using the normal distribution.

- $Ha: p < p_0 \Rightarrow p-value = P(Z \leq z)$:

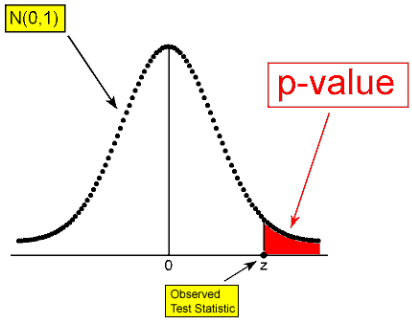


Looking at the shaded region, you can see why this is often referred to as a **left-tailed** test. We shaded to the left of the test statistic, since less than is to the left.

Alternative Hypothesis is "Greater Than"

The probability of observing a test statistic as **large as that observed or larger**, assuming that the values of the test statistic follow a standard normal distribution. Again, we will represent this probability in symbols and using the normal distribution

- $Ha: p > p_0 \Rightarrow p-value = P(Z \geq z)$:

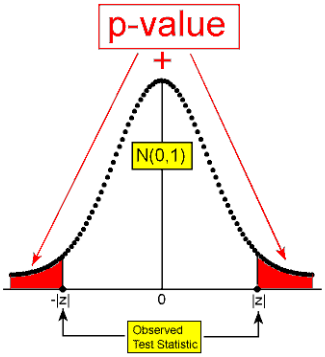


Looking at the shaded region, you can see why this is often referred to as a **right-tailed** test. We shaded to the right of the test statistic, since greater than is to the right.

Alternative Hypothesis is "Not Equal To"

The probability of observing a test statistic which is as large in **magnitude** as that observed or larger, assuming that the values of the test statistic follow a standard normal distribution.

- $Ha: p \neq p_0 \Rightarrow p-value = P(Z \leq -|z|) + P(Z \geq |z|) = 2P(Z \geq |z|)$:



This is often referred to as a **two-tailed** test, since we shaded in both directions.

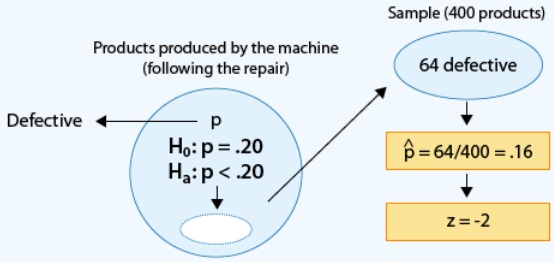
Next, we will apply this to our three examples. But first, work through the following activities, which should help your understanding.

Learn by Doing: Proportions (Step 3)

Did I Get This?: Proportions (Step 3)

EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?



The p-value in this case is:

- The probability of observing a test statistic as small as -2 or smaller, assuming that Ho is true.

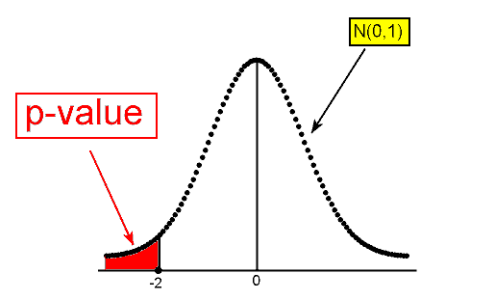
OR (recalling what the test statistic actually means in this case),

- The probability of observing a sample proportion that is 2 standard deviations or more below the null value ($p_0 = 0.20$), assuming that p_0 is the true population proportion.

OR, more specifically,

- The probability of observing a sample proportion of 0.16 or lower in a random sample of size 400, when the true population proportion is $p_0 = 0.20$

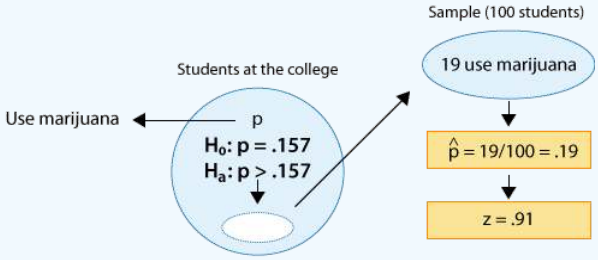
In either case, the p-value is found as shown in the following figure:



To find $P(Z \leq -2)$ we can either use the calculator or table we learned to use in the probability unit for normal random variables. Eventually, after we understand the details, we will use software to run the test for us and the output will give us all the information we need. The p-value that the statistical software provides for this specific example is 0.023. The p-value tells us that it is pretty unlikely (probability of 0.023) to get data like those observed (test statistic of -2 or less) assuming that Ho is true.

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?



The p-value in this case is:

- The probability of observing a test statistic as large as 0.91 or larger, assuming that Ho is true.

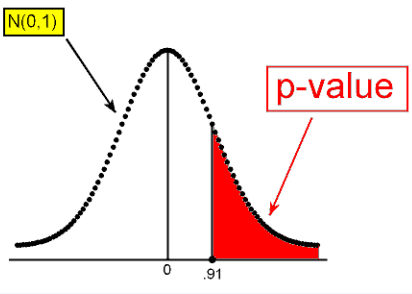
OR (recalling what the test statistic actually means in this case),

- The probability of observing a sample proportion that is 0.91 standard deviations or more above the null value ($p_0 = 0.157$), assuming that p_0 is the true population proportion.

OR, more specifically,

- The probability of observing a sample proportion of 0.19 or higher in a random sample of size 100, when the true population proportion is $p_0=0.157$

In either case, the p-value is found as shown in the following figure:



Again, at this point we can either use the calculator or table to find that the p-value is 0.182, this is P(Z ≥ 0.91).

The p-value tells us that it is not very surprising (probability of 0.182) to get data like those observed (which yield a test statistic of 0.91 or higher) assuming that the null hypothesis is true.

EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?

The p-value in this case is:

- The probability of observing a test statistic as large as 2.31 (or larger) or as small as -2.31 (or smaller), assuming that Ho is true.

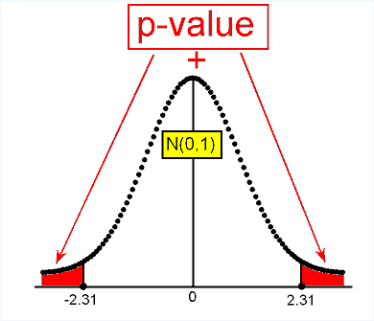
OR (recalling what the test statistic actually means in this case),

- The probability of observing a sample proportion that is 2.31 standard deviations or more away from the null value (p_0 = 0.64), assuming that $p_0$is the true population proportion.

OR, more specifically,

- The probability of observing a sample proportion as different as 0.675 is from 0.64, or even more different (i.e. as high as 0.675 or higher or as low as 0.605 or lower) in a random sample of size 1,000, when the true population proportion is p_0= 0.64

In either case, the p-value is found as shown in the following figure:



Again, at this point we can either use the calculator or table to find that the p-value is 0.021, this is P(Z ≤ -2.31) + P(Z ≥ 2.31) = 2*P(Z ≥ |2.31|)

The p-value tells us that it is pretty unlikely (probability of 0.021) to get data like those observed (test statistic as high as 2.31 or higher or as low as -2.31 or lower) assuming that Ho is true.

Comment:

- We've just seen that finding p-values involves probability calculations about the value of the test statistic assuming that Ho is true. In this case, when Ho is true, the values of the test statistic follow a standard normal distribution (i.e., the sampling distribution of the test statistic when the null hypothesis is true is N(0,1)). Therefore, p-values correspond to areas (probabilities) under the standard normal curve.

Similarly, in **any test**, p-values are found using the sampling distribution of the test statistic when the null hypothesis is true (also known as the "null distribution" of the test statistic). In this case, it was relatively easy to argue that the null distribution of our test statistic is N(0,1). As we'll see, in other tests, other distributions come up (like the t-distribution and the F-distribution), which we will just mention briefly, and rely heavily on the output of our statistical package for obtaining the p-values.

We've just completed our discussion about the p-value, and how it is calculated both in general and more specifically for the z-test for the population proportion. Let's go back to the four-step process of hypothesis testing and see what we've covered and what still needs to be discussed.

The Four Steps in Hypothesis Testing

- **STEP 1:** State the appropriate null and alternative hypotheses, Ho and Ha.
- **STEP 2:** Obtain a random sample, collect relevant data, and **check whether the data meet the conditions under which the test can be used**. If the conditions are met, summarize the data using a test statistic.
- **STEP 3:** Find the p-value of the test.
- **STEP 4:** Based on the p-value, decide whether or not the results are statistically significant and **draw your conclusions in context.**
- **Note:** In practice, we should always consider the practical significance of the results as well as the statistical significance.

With respect to the z-test the population proportion:

Step 1: Completed

Step 2: Completed

Step 3: Completed

Step 4. This is what we will work on next.

Learn by Doing: Proportions (Step 3) Understanding P-values

Proportions (Step 4 & Summary)

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.33: In a given context, distinguish between situations involving a population proportion and a population mean and specify the correct null and alternative hypothesis for the scenario.

Learning Objectives

LO 4.34: Carry out a complete hypothesis test for a population proportion by hand.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Learning Objectives

LO 6.27: Explain what the p-value is and how it is used to draw conclusions.

Video

Video: Proportions (Step 4 & Summary) (4:30)

Step 4. Drawing Conclusions Based on the P-Value

This last part of the four-step process of hypothesis testing is the same across all statistical tests, and actually, we've already said basically everything there is to say about it, but it can't hurt to say it again.

The p-value is a measure of how much evidence the data present against Ho. The smaller the p-value, the more evidence the data present against Ho.

We already mentioned that what determines what constitutes enough evidence against Ho is the significance level (α, alpha), a cutoff point below which the p-value is considered small enough to reject Ho in favor of Ha. The most commonly used significance level is 0.05.

- If p-value ≤ 0.05 then **WE REJECT** Ho
  - Conclusion: There **IS** enough evidence that *Ha is True*
- If p-value > 0.05 then **WE FAIL TO REJECT** Ho
  - Conclusion: There **IS NOT** enough evidence that *Ha is True*

Where instead of *Ha is True*, we write what this means in the words of the problem, in other words, in the context of the current scenario.

It is important to mention again that this step has essentially two sub-steps:

- (i) Based on the p-value, determine whether or not the results are statistically significant (i.e., the data present enough evidence to reject Ho).
- (ii) State your conclusions in the context of the problem.

Note: We always still must consider whether the results have any practical significance, particularly if they are statistically significant as a statistically significant result which has not practical use is essentially meaningless!

Let's go back to our three examples and draw conclusions.

EXAMPLE:

Has the proportion of defective products been reduced as a result of the repair?

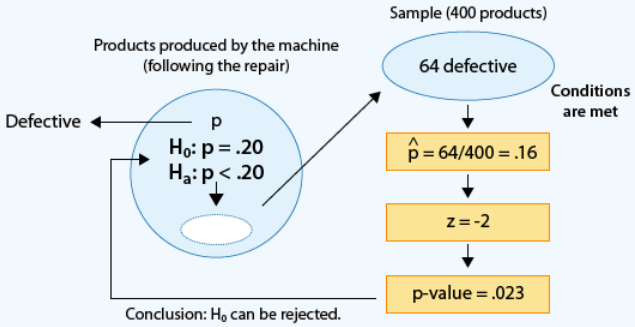
We found that the p-value for this test was 0.023.

Since 0.023 is small (in particular, 0.023 < 0.05), the data provide enough evidence to reject Ho.

Conclusion:

- There **IS** enough evidence that *the proportion of defective products is less than 20% after the repair*.

The following figure is the complete story of this example, and includes all the steps we went through, starting from stating the hypotheses and ending with our conclusions:



EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?

We found that the p-value for this test was 0.182.

Since .182 is *not* small (in particular, 0.182 > 0.05), the data do not provide enough evidence to reject Ho.

Conclusion:

- There **IS NOT** enough evidence that *the proportion of students at the college who use marijuana is higher than the national figure.*

Here is the complete story of this example:

A large circle represents the population Students at the college. We want to know p about this population, or what is the population proportion of students using marijuana. The hypotheses are H_0: p = .157 and H_a: p .157 . We take a sample of 100 students, represented by a smaller circle. We find that 19 use marijuana. p-hat = 19/100 = .19, z = .91, and p-value = .182 . Since the p-value is too large we conclude that H_0 cannot be rejected." height="278" loading="lazy" src="http://phhp-faculty-cantrell.sites.m...3/image276.gif" title="A large circle represents the population Students at the college. We want to know p about this population, or what is the population proportion of students using marijuana. The hypotheses are H_0: p = .157 and H_a: p > .157 . We take a sample of 100 students, represented by a smaller circle. We find that 19 use marijuana. p-hat = 19/100 = .19, z = .91, and p-value = .182 . Since the p-value is too large we conclude that H_0 cannot be rejected." width="564">

Learn by Doing: Learn by Doing – Proportions (Step 4)

EXAMPLE:

Did the proportion of U.S. adults who support the death penalty change between 2003 and a later poll?

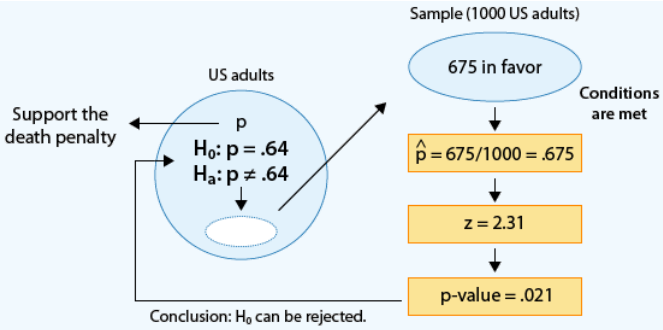
We found that the p-value for this test was 0.021.

Since 0.021 is small (in particular, 0.021 < 0.05), the data provide enough evidence to reject Ho

Conclusion:

- There **IS** enough evidence that *the proportion of adults who support the death penalty for convicted murderers has changed since 2003.*

Here is the complete story of this example:



Did I Get This?: Proportions (Step 4)

Many Students Wonder: **Hypothesis Testing for the Population Proportion**

Many students wonder why 5% is often selected as the significance level in hypothesis testing, and why 1% is the next most typical level. This is largely due to just convenience and tradition.

When Ronald Fisher (one of the founders of modern statistics) published one of his tables, he used a mathematically convenient scale that included 5% and 1%. Later, these same 5% and 1% levels were used by other people, in part just because Fisher was so highly esteemed. But mostly these are arbitrary levels.

The idea of selecting some sort of relatively small cutoff was historically important in the development of statistics; but it's important to remember that there is really a continuous range of increasing confidence towards the alternative hypothesis, not a single all-or-nothing value. There isn't much meaningful difference, for instance, between a p-value of .049 or .051, and it would be foolish to declare one case definitely a "real" effect and to declare the other case definitely a "random" effect. In either case, the study results were roughly 5% likely by chance if there's no actual effect.

Whether such a p-value is sufficient for us to reject a particular null hypothesis ultimately depends on the risk of making the wrong decision, and the extent to which the hypothesized effect might contradict our prior experience or previous studies.

Let's Summarize!!

We have now completed going through the four steps of hypothesis testing, and in particular we learned how they are applied to the z-test for the population proportion. Here is a brief summary:

- **Step 1: State the hypotheses**

  State the null hypothesis:

  Ho: $p = p_0$

  State the alternative hypothesis:

  Ha: $p < p_0$ **(one-sided)**

  Ha: $p > p_0$ **(one-sided)**

  Ha: $p \neq p_0$ **(two-sided)**

  where the choice of the appropriate alternative (out of the three) is usually quite clear from the context of the problem. If you feel it is not clear, it is most likely a two-sided problem. Students are usually good at recognizing the "more than" and "less than" terminology but differences can sometimes be more difficult to spot, sometimes this is because you have preconceived ideas of how you think it should be! Use only the information given in the problem.

- **Step 2: Obtain data, check conditions, and summarize data**

  Obtain data from a sample and:

  (i) Check whether the data satisfy the conditions which allow you to use this test.

random sample (or at least a sample that can be considered random in context)

the conditions under which the sampling distribution of p-hat is normal are met

(ii) Calculate the sample proportion p-hat, and summarize the data using the test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

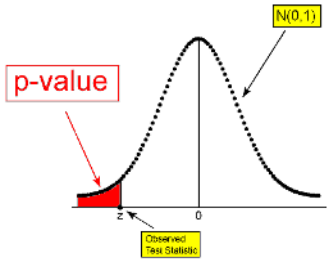
(**Recall:** This standardized test statistic represents how many standard deviations above or below p_0 our sample proportion p-hat is.)

- **Step 3: Find the p-value of the test by using the test statistic as follows**

IMPORTANT FACT: In all future tests, we will rely on software to obtain the p-value.

When the alternative hypothesis is "less than" the probability of observing a test statistic as **small as that observed or smaller**, assuming that the values of the test statistic follow a standard normal distribution. We will now represent this probability in symbols and also using the normal distribution.

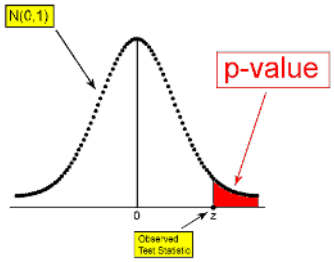
- $Ha: p < p_0 \Rightarrow p-value = P(Z \leq z)$:



Looking at the shaded region, you can see why this is often referred to as a **left-tailed** test. We shaded to the left of the test statistic, since less than is to the left.

When the alternative hypothesis is "greater than" the probability of observing a test statistic as **large as that observed or larger**, assuming that the values of the test statistic follow a standard normal distribution. Again, we will represent this probability in symbols and using the normal distribution

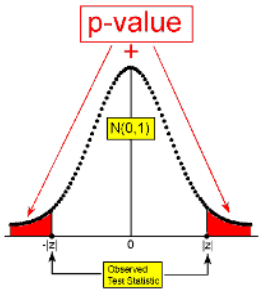
- $Ha: p > p_0 \Rightarrow p-value = P(Z \geq z)$:



Looking at the shaded region, you can see why this is often referred to as a **right-tailed** test. We shaded to the right of the test statistic, since greater than is to the right.

When the alternative hypothesis is "not equal to" the probability of observing a test statistic which is as large in **magnitude** as that observed or larger, assuming that the values of the test statistic follow a standard normal distribution.

- $Ha: p \neq p_0 \Rightarrow p-value = P(Z \leq -|z|)+P(Z \geq |z|) = 2P(Z \geq |z|)$:



This is often referred to as a **two-tailed** test, since we shaded in both directions.

- **Step 4: Conclusion**

  Reach a conclusion first regarding the statistical significance of the results, and then determine what it means in the context of the problem.

  If p-value ≤ 0.05 then WE REJECT Ho
  Conclusion: There IS enough evidence that Ha is True

  If p-value > 0.05 then WE FAIL TO REJECT Ho
  Conclusion: There IS NOT enough evidence that Ha is True

  Recall that: If the p-value is small (in particular, smaller than the significance level, which is usually 0.05), the results are statistically significant (in the sense that there is a statistically significant difference between what was observed in the sample and what was claimed in Ho), and so we reject Ho.

  If the p-value is not small, we do not have enough statistical evidence to reject Ho, and so we continue to believe that Ho **may** be true. (**Remember: In hypothesis testing we never "accept" Ho**).

  Finally, in practice, we should always consider the **practical significance** of the results as well as the statistical significance.

Learn by Doing: Z-Test for a Population Proportion

What's next?

Before we move on to the next test, we are going to use the z-test for proportions to bring up and illustrate a few more very important issues regarding hypothesis testing. This might also be a good time to review the concepts of Type I error, Type II error, and Power before continuing on.

More about Hypothesis Testing

CO-1: Describe the roles biostatistics serves in the discipline of public health.

Learning Objectives

LO 1.11: Recognize the distinction between statistical significance and practical significance.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Learning Objectives

LO 6.30: Use a confidence interval to determine the correct conclusion to the associated two-sided hypothesis test.

Video

Video: More about Hypothesis Testing (18:25)

The issues regarding hypothesis testing that we will discuss are:

1. The effect of sample size on hypothesis testing.
2. Statistical significance vs. practical importance.
3. Hypothesis testing and confidence intervals—how are they related?

Let's begin.

1. The Effect of Sample Size on Hypothesis Testing

We have already seen the effect that the sample size has on inference, when we discussed point and interval estimation for the population mean (μ, mu) and population proportion (p). Intuitively …

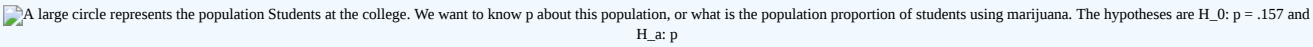
Larger sample sizes give us more information to pin down the true nature of the population. We can therefore expect the **sample** mean and **sample** proportion obtained from a larger sample to be closer to the population mean and proportion, respectively. As a result, for the same level of confidence, we can report a smaller margin of error, and get a narrower confidence interval. What we've seen, then, is that larger sample size gives a boost to how much we trust our sample results.

In hypothesis testing, larger sample sizes have a similar effect. We have also discussed that the power of our test increases when the sample size increases, all else remaining the same. This means, we have a better chance to detect the difference between the true value and the null value for larger samples.

The following two examples will illustrate that a larger sample size provides more convincing evidence (the test has greater power), and how the evidence manifests itself in hypothesis testing. Let's go back to our example 2 (marijuana use at a certain liberal arts college).

EXAMPLE:

Is the proportion of marijuana users in the college higher than the national figure?

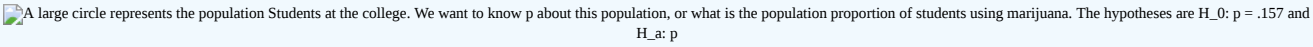


.157 . We take a sample of 100 students, represented by a smaller circle. We find that 19 use marijuana. p-hat = 19/100 = .19, z = .91, and p-value = .182 . Since the p-value is too large we conclude that H_0 cannot be rejected." height="278" loading="lazy" src="http://phhp-faculty-cantrell.sites.m...3/image276.gif" title="A large circle represents the population Students at the college. We want to know p about this population, or what is the population proportion of students using marijuana. The hypotheses are H_0: p = .157 and H_a: p > .157 . We take a sample of 100 students, represented by a smaller circle. We find that 19 use marijuana. p-hat = 19/100 = .19, z = .91, and p-value = .182 . Since the p-value is too large we conclude that H_0 cannot be rejected." width="564">

We do **not** have enough evidence to conclude that the proportion of students at the college who use marijuana is higher than the national figure.

Now, let's increase the sample size.

There are rumors that students in a certain liberal arts college are more inclined to use drugs than U.S. college students in general. Suppose that **in a simple random sample of 400 students from the college, 76 admitted to marijuana use**. Do the data provide enough evidence to conclude that the proportion of marijuana users among the students in the college (p) is **higher** than the national proportion, which is 0.157? (Reported by the Harvard School of Public Health).



.157 . We take a sample of 400 students, represented by a smaller circle, and find that 76 use marijuana. Conditions are met to

use our method, so p-hat = 76/400 = .19, z = 1.81, and p-value = .035 . The p-value is low enough to let us conclude that we can reject H_0." height="292" loading="lazy" src="http://phhp-faculty-cantrell.sites.m...3/image291.gif" title="A large circle represents the population Students at the college. We want to know p about this population, or what is the population proportion of students using marijuana. The hypotheses are H_0: p = .157 and H_a: p > .157 . We take a sample of 400 students, represented by a smaller circle, and find that 76 use marijuana. Conditions are met to use our method, so p-hat = 76/400 = .19, z = 1.81, and p-value = .035 . The p-value is low enough to let us conclude that we can reject H_0." width="572">

Our results here are statistically **significant**. In other words, in example 2* the data provide enough evidence to reject Ho.

- **Conclusion:** There is enough evidence that the proportion of marijuana users at the college is higher than among all U.S. students.

What do we learn from this?

We see that sample results that are based on a larger sample carry more weight (have greater power).

In example 2, we saw that a sample proportion of 0.19 based on a sample of size of 100 was not enough evidence that the proportion of marijuana users in the college is higher than 0.157. Recall, from our general overview of hypothesis testing, that this conclusion (not having enough evidence to reject the null hypothesis) **doesn't** mean the null hypothesis is necessarily true (so, we never "accept" the null); it only means that the particular study didn't yield sufficient evidence to reject the null. It **might** be that the sample size was simply too small to detect a statistically significant difference.

However, in example 2*, we saw that when the sample proportion of 0.19 is obtained from a sample of size 400, it carries much more weight, and in particular, provides enough evidence that the proportion of marijuana users in the college is higher than 0.157 (the national figure). In **this** case, the sample size of 400 **was** large enough to detect a statistically significant difference.

The following activity will allow you to practice the ideas and terminology used in hypothesis testing when a result is not statistically significant.

Learn by Doing: Interpreting Non-significant Results

2. Statistical significance vs. practical importance.

Now, we will address the issue of statistical significance versus practical importance (which also involves issues of sample size).

The following activity will let you explore the effect of the sample size on the statistical significance of the results yourself, and more importantly will discuss issue **2: Statistical significance vs. practical importance.**

Important Fact: In general, with a sufficiently large sample size you can make any result that has very little practical importance statistically significant! A large sample size alone does NOT make a "good" study!!

This suggests that when interpreting the results of a test, you should always think not only about the statistical significance of the results but also about their practical importance.

Learn by Doing: Statistical vs. Practical Significance

3. Hypothesis Testing and Confidence Intervals

The last topic we want to discuss is the relationship between hypothesis testing and confidence intervals. Even though the flavor of these two forms of inference is different (confidence intervals estimate a parameter, and hypothesis testing assesses the evidence in the data against one claim and in favor of another), there is a strong link between them.

We will explain this link (using the z-test and confidence interval for the population proportion), and then explain how confidence intervals can be used after a test has been carried out.

Recall that a confidence interval gives us a set of plausible values for the unknown population parameter. We may therefore examine a confidence interval to informally decide if a proposed value of population proportion seems plausible.

For example, if a 95% confidence interval for p, the proportion of all U.S. adults already familiar with Viagra in May 1998, was (0.61, 0.67), then it seems clear that we should be able to reject a claim that only 50% of all U.S. adults were familiar with the drug, since based on the confidence interval, 0.50 is not one of the plausible values for p.

In fact, the information provided by a confidence interval can be formally related to the information provided by a hypothesis test. (**Comment:** The relationship is more straightforward for two-sided alternatives, and so we will not present results for the one-sided cases.)

Suppose we want to carry out the **two-sided test:**

- Ho: $p = p_0$
- Ha: $p \neq p_0$

using a significance level of 0.05.

An alternative way to perform this test is to find a 95% **confidence interval** for p and check:

- If p_0 falls **outside** the confidence interval, **reject** Ho.
- If p_0 falls **inside** the confidence interval, **do not reject** Ho.

In other words,

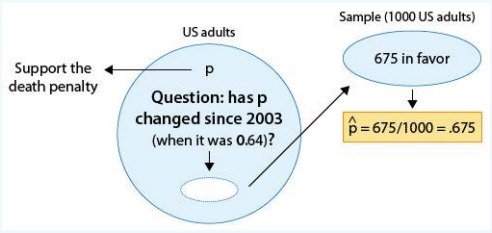
- If p_0 is not one of the plausible values for p, we reject Ho.
- If p_0 is a plausible value for p, we cannot reject Ho.

(**Comment:** Similarly, the results of a test using a significance level of 0.01 can be related to the 99% confidence interval.)

Let's look at an example:

EXAMPLE:

Recall example 3, where we wanted to know whether the proportion of U.S. adults who support the death penalty for convicted murderers has changed since 2003, when it was 0.64.



We are testing:

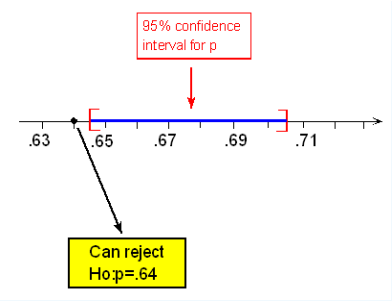
- **Ho:** p = 0.64 (No change from 2003).
- **Ha:** p ≠ 0.64 (Some change since 2003).

and as the figure reminds us, we took a sample of 1,000 U.S. adults, and the data told us that 675 supported the death penalty for convicted murderers (p-hat = 0.675).

A 95% confidence interval for p, the proportion of **all** U.S. adults who support the death penalty, is:

$$0.675 \pm 1.96\sqrt{\frac{0.675(1-0.675)}{1000}} \approx 0.675 \pm 0.029 = (0.646, 0.704)$$

Since the 95% confidence interval for p does not include 0.64 as a plausible value for p, we can reject Ho and conclude (as we did before) that there is enough evidence that the proportion of U.S. adults who support the death penalty for convicted murderers has changed since 2003.



EXAMPLE:

You and your roommate are arguing about whose turn it is to clean the apartment. Your roommate suggests that you settle this by tossing a coin and takes one out of a locked box he has on the shelf. Suspecting that the coin might not be fair, you decide to test it first. You toss the coin 80 times, thinking to yourself that if, indeed, the coin is fair, you should get around 40 heads. Instead you get 48 heads. You are puzzled. You are not sure whether getting 48 heads out of 80 is enough evidence to conclude that the coin is unbalanced, or whether this a result that could have happened just by chance when the coin is fair.

Statistics can help you answer this question.

Let p be the true proportion (probability) of heads. We want to test whether the coin is fair or not.

We are testing:

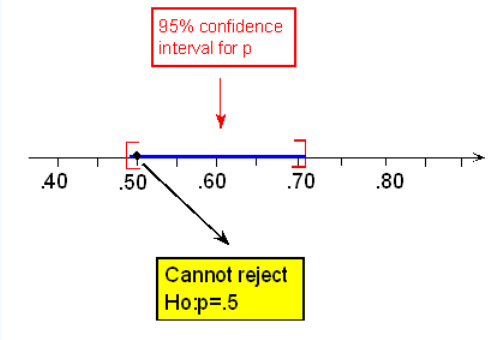
- **Ho:** p = 0.5 (the coin is fair).
- **Ha:** p ≠ 0.5 (the coin is not fair).

The data we have are that out of n = 80 tosses, we got 48 heads, or that the sample proportion of heads is p-hat = 48/80 = 0.6.

A 95% confidence interval for p, the true proportion of heads for this coin, is:

$$0.6 \pm 1.96\sqrt{\frac{0.6(1-0.6)}{80}} \approx 0.6 \pm 0.11 = (0.49, 0.71)$$

Since in this case 0.5 is one of the plausible values for p, we cannot reject Ho. In other words, the data do not provide enough evidence to conclude that the coin is not fair.



Comment

The context of the last example is a good opportunity to bring up an important point that was discussed earlier.

Even though we use 0.05 as a cutoff to guide our decision about whether the results are statistically significant, we should not treat it as inviolable and we should always add our own judgment. Let's look at the last example again.

It turns out that the p-value of this test is 0.0734. In other words, it is maybe not extremely unlikely, but it is quite unlikely (probability of 0.0734) that when you toss a fair coin 80 times you'll get a sample proportion of heads of 48/80 = 0.6 (or even more extreme). It is true that using the 0.05 significance level (cutoff), 0.0734 is not considered small enough to conclude that the coin is not fair. However, if you really don't want to clean the apartment, the p-value might be small enough for you to ask your roommate to use a different coin, or to provide one yourself!

Did I Get This?: Connection between Confidence Intervals and Hypothesis Tests

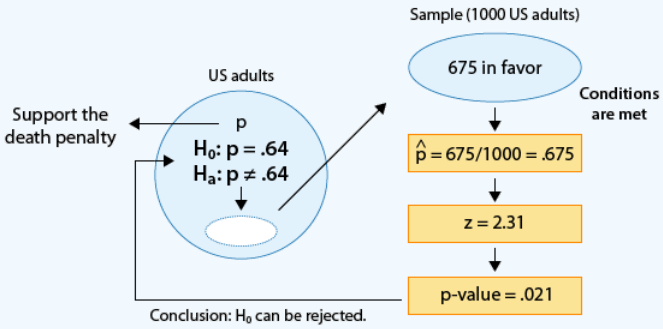
Did I Get This?: Hypothesis Tests for Proportions (Extra Practice)

Here is our final point on this subject:

When the data provide enough evidence to reject Ho, we can conclude (depending on the alternative hypothesis) that the population proportion is either less than, greater than, or not equal to the null value p_0. However, we do not get a more informative statement about its actual value. It might be of interest, then, to follow the test with a 95% confidence interval that will give us more insight into the actual value of p.

✔ EXAMPLE:

In our example 3,



we concluded that the proportion of U.S. adults who support the death penalty for convicted murderers has changed since 2003, when it was 0.64. It is probably of interest not only to know that the proportion has changed, but also to estimate what it has changed to. We've calculated the 95% confidence interval for p on the previous page and found that it is (0.646, 0.704).

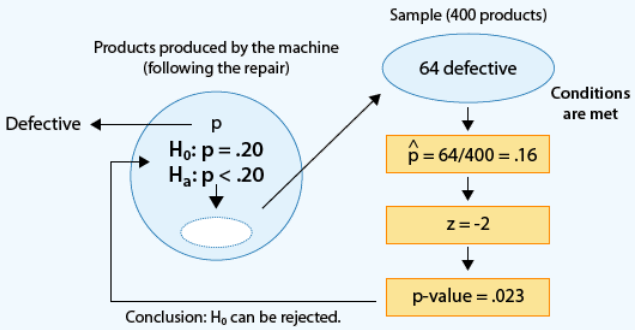
We can combine our conclusions from the test and the confidence interval and say:

Data provide evidence that the proportion of U.S. adults who support the death penalty for convicted murderers has changed since 2003, and we are 95% confident that it is now between 0.646 and 0.704. (i.e. between 64.6% and 70.4%).

✔ EXAMPLE:

Let's look at our example 1 to see how a confidence interval following a test might be insightful in a different way.

Here is a summary of example 1:



We conclude that as a result of the repair, the proportion of defective products has been reduced to below 0.20 (which was the proportion prior to the repair). It is probably of great interest to the company not only to know that the proportion of defective has been reduced, but also estimate what it has been reduced to, to get a better sense of how effective the repair was. A 95% confidence interval for p in this case is:

$$0.16 \pm 1.96\sqrt{\frac{0.16(1-0.16)}{400}} \approx 0.16 \pm 0.036 = (0.124, 0.196)$$

We can therefore say that the data provide evidence that the proportion of defective products has been reduced, and we are 95% confident that it has been reduced to somewhere between 12.4% and 19.6%. This is very useful information, since it tells us that even though the results were significant (i.e., the repair reduced the number of defective products), the repair might not have been effective enough, if it managed to reduce the number of defective products only to the range provided by the confidence interval. This, of course, ties back in to the idea of statistical significance vs. practical importance that we discussed earlier. Even though the results are statistically significant (Ho was rejected), practically speaking, the repair might still be considered ineffective.

Learn by Doing: Hypothesis Tests and Confidence Intervals

Let's summarize

Even though this portion of the current section is about the z-test for population proportion, it is loaded with very important ideas that apply to hypothesis testing in general. We've already summarized the details that are specific to the z-test for proportions, so the purpose of this summary is to highlight the general ideas.

The process of hypothesis testing has **four steps**:

I. Stating the null and alternative hypotheses (Ho and Ha).

II. Obtaining a random sample (or at least one that can be considered random) and collecting data. Using the data:

Check that the conditions under which the test can be reliably used are met.

Summarize the data using a test statistic.

- The test statistic is a measure of the evidence in the data against Ho. The larger the test statistic is in magnitude, the more evidence the data present against Ho.

III. Finding the p-value of the test. The p-value is the probability of getting data like those observed (or even more extreme) assuming that the null hypothesis is true, and is calculated using the null distribution of the test statistic. The p-value is a measure of the evidence against Ho. The smaller the p-value, the more evidence the data present against Ho.

IV. Making conclusions.

Conclusions about the statistical **significance of the results:**

If the p-value is small, the data present enough evidence to reject Ho (and accept Ha).

If the p-value is not small, the data do not provide enough evidence to reject Ho.

To help guide our decision, we use the significance level as a cutoff for what is considered a small p-value. The significance cutoff is usually set at 0.05.

Conclusions should then be provided **in the context** of the problem.

Additional Important Ideas about Hypothesis Testing

- Results that are based on a larger sample carry more weight, and therefore **as the sample size increases, results become more statistically significant.**
- Even a very small and practically unimportant effect becomes statistically significant with a large enough sample size. The **distinction between statistical significance and practical importance** should therefore always be considered.
- **Confidence intervals can be used in order to carry out two-sided tests** (95% confidence for the 0.05 significance level). If the null value is not included in the confidence interval (i.e., is not one of the plausible values for the parameter), we have enough evidence to reject Ho. Otherwise, we cannot reject Ho.
- If the results are statistically significant, it might be of interest to **follow up the tests with a confidence interval** in order to get insight into the actual value of the parameter of interest.

- It is important to be aware that there are two types of errors in hypothesis testing (**Type I and Type II**) and that the **power** of a statistical test is an important measure of how likely we are to be able to detect a difference of interest to us in a particular problem.

Means (All Steps)

NOTE: Beginning on this page, the Learn By Doing and Did I Get This activities are presented as interactive PDF files. The interactivity may not work on mobile devices or with certain PDF viewers. Use an official ADOBE product such as ADOBE READER.

If you have any issues with the **Learn By Doing** or **Did I Get This** interactive PDF files, you can view all of the questions and answers presented on this page in this document:

- QUESTION/Answer (SPOILER ALERT!)

Learning Objectives

LO 4.33: In a given context, distinguish between situations involving a population proportion and a population mean and specify the correct null and alternative hypothesis for the scenario.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Learning Objectives

LO 6.26: Outline the logic and process of hypothesis testing.

Learning Objectives

LO 6.27: Explain what the p-value is and how it is used to draw conclusions.

Learning Objectives

LO 6.30: Use a confidence interval to determine the correct conclusion to the associated two-sided hypothesis test.

Video

Video: Means (All Steps) (13:11)

So far we have talked about the logic behind hypothesis testing and then illustrated how this process proceeds in practice, using the z-test for the population proportion (p).

We are now moving on to discuss **testing for the population mean (μ, mu),** which is the parameter of interest when the variable of interest is quantitative.

A few comments about the structure of this section:

- The **basic groundwork for carrying out hypothesis tests** has already been laid in our general discussion and in our presentation of tests about proportions.

Therefore we can easily modify the four steps to carry out tests about means instead, without going into all of the details again.

We will use this approach for all future tests so **be sure to go back to the discussion in general and for proportions to review the concepts in more detail.**

- In our discussion about confidence intervals for the population mean, we made the distinction between whether the population standard deviation, σ (sigma) was known or if we needed to estimate this value using the sample standard deviation, *s*.

In this section, we will only discuss the second case as in most realistic settings we do not know the population standard deviation.

In this case we need to use the *t*-distribution instead of the standard normal distribution for the probability aspects of confidence intervals (choosing table values) and hypothesis tests (finding p-values).

- Although we will discuss some theoretical or conceptual details for some of the analyses we will learn, **from this point on we will rely on software to conduct tests and calculate confidence intervals for us**, while we **focus on understanding which methods are used for which situations and what the results say in context.**

If you are interested in more information about the z-test, where we assume the population standard deviation σ (sigma) is known, you can review the Carnegie Mellon Open Learning Statistics Course (you will need to click "ENTER COURSE").

Like any other tests, the ***t*-test for the population mean follows the four-step process:**

- **STEP 1:** Stating the hypotheses $H_o$and H_a.
- **STEP 2:** Collecting relevant data, checking that the data satisfy the conditions which allow us to use this test, and summarizing the data using a test statistic.
- **STEP 3:** Finding the p-value of the test, the probability of obtaining data as extreme as those collected (or even more extreme, in the direction of the alternative hypothesis), assuming that the null hypothesis is true. In other words, how likely is it that the only reason for getting data like those observed is sampling variability (and not because $H_o$is not true)?
- **STEP 4:** Drawing conclusions, assessing the statistical significance of the results based on the p-value, and stating our conclusions in context. (Do we or don't we have evidence to reject $H_o$and accept H_a?)
- **Note:** In practice, we should also always consider the practical significance of the results as well as the statistical significance.

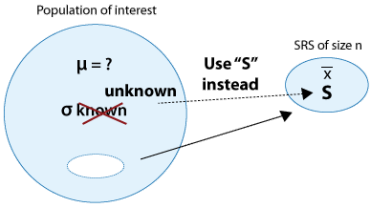
We will now go through the four steps specifically for the *t*-test for the population mean and apply them to our two examples.

Tests About μ (mu) When σ (sigma) is Unknown – The t-test for a Population Mean

Only in a few cases is it reasonable to assume that the population standard deviation, σ (sigma), is known and so we will not cover hypothesis tests in this case. We discussed both cases for confidence intervals so that we could still calculate some confidence intervals by hand.

For this and all future tests we will rely on software to obtain our summary statistics, test statistics, and p-values for us.

The case where σ (sigma) is unknown is much more common in practice. What can we use to replace σ (sigma)? If you don't know the population standard deviation, the best you can do is find the sample standard deviation, s, and use it instead of σ (sigma). (Note that this is exactly what we did when we discussed confidence intervals).



Is that it? Can we just use s instead of σ (sigma), and the rest is the same as the previous case? Unfortunately, it's not that simple, but not very complicated either.

Here, when we use the sample standard deviation, s, as our estimate of σ (sigma) we can no longer use a normal distribution to find the cutoff for confidence intervals or the p-values for hypothesis tests.

> **Instead we must use the *t*-distribution (with n-1 degrees of freedom) to obtain the p-value for this test.**

We discussed this issue for confidence intervals. We will talk more about the *t*-distribution after we discuss the details of this test for those who are interested in learning more.

It isn't really necessary for us to understand this distribution but it is important that we use the correct distributions in practice via our software.

We will wait until UNIT 4B to look at how to accomplish this test in the software. For now focus on understanding the process and drawing the correct conclusions from the p-values given.

Now let's go through the four steps in conducting the *t*-test for the population mean.

Step 1: State the hypotheses

The null and alternative hypotheses for the *t*-test for the population mean (μ, mu) have exactly the same structure as the hypotheses for z-test for the population proportion (p):

> The **null hypothesis** has the form:
>
> - Ho: $\mu = \mu_0$ (mu = mu_zero)
>
> (where μ_0 (mu_zero) is often called the null value)
>
> The **alternative hypothesis** takes one of the following three forms (depending on the context):
>
> - Ha: $\mu < \mu_0$ (mu < mu_zero) **(one-sided)**
> - Ha: $\mu > \mu_0$ (mu > mu_zero) **(one-sided)**
> - Ha: $\mu \neq \mu_0$ (mu ≠ mu_zero) **(two-sided)**
>
> where the choice of the appropriate alternative (out of the three) is usually quite clear from the context of the problem.
>
> **If you feel it is not clear, it is most likely a two-sided problem.** Students are usually good at recognizing the "more than" and "less than" terminology but differences can sometimes be more difficult to spot, sometimes this is because you have preconceived ideas of how you think it should be! You also cannot use the information from the sample to help you determine the hypothesis. We would not know our data when we originally asked the question.

Now try it yourself. Here are a few exercises on stating the hypotheses for tests for a population mean.

> **Learn by Doing:** State the Hypotheses for a test for a population mean

Here are a few more activities for practice.

> **Did I Get This?:** State the Hypotheses for a test for a population mean

When setting up hypotheses, be sure to use only the information in the research question. We cannot use our sample data to help us set up our hypotheses.

For this test, it is still important to correctly choose the alternative hypothesis as "less than", "greater than", or "different" although generally in practice two-sample tests are used.

Step 2: Obtain data, check conditions, and summarize data

Obtain data from a sample:

- In this step we would **obtain data from a sample.** This is not something we do much of in courses but it is done very often in practice!

Check the conditions:

- Then we **check the conditions under which this test (the *t*-test for one population mean) can be safely carried out** – which are:

> - The **sample is random** (or at least can be considered random in context).
> - **We are in one of the three situations marked with a green check mark** in the following table (which ensure that x-bar is at least approximately normal and the test statistic using the sample standard deviation, s, is therefore a *t*-distribution with n-1 degrees of freedom – proving this is beyond the scope of this course):
>
> | **Conditions: t-test for a population mean** | Small sample size | Large sample size |
> |---|---|---|
> | Variable varies normally in the population | ✓ | ✓ |
> | Variable doesn't vary normally in the population | ✗ | ✓ |
>
> - **For large samples,** we **don't need to check for normality in the population**. We can rely on the sample size as the basis for the validity of using this test.
> - **For small samples**, we **need to have data from a normal population** in order for the p-values and confidence intervals to be valid.

In practice, for small samples, it can be very difficult to determine if the population is normal. Here is a simulation to give you a better understanding of the difficulties.

> **Video:** Simulations – Are Samples from a Normal Population? (4:58)

Now try it yourself with a few activities.

> **Learn by Doing:** Checking Conditions for Hypothesis Testing for the Population Mean

Comments:

- It is always a good idea to look at the data and get a sense of their pattern regardless of whether you actually need to do it in order to assess whether the conditions are met.
- This idea of looking at the data is relevant to all tests in general. In the next module—inference for relationships—conducting exploratory data analysis before inference will be an integral part of the process.

Here are a few more problems for extra practice.

> **Did I Get This?:** Checking Conditions for Hypothesis Testing for the Population Mean

When setting up hypotheses, be sure to use only the information in the res

Calculate Test Statistic

Assuming that the conditions are met, we calculate the sample mean x-bar and the sample standard deviation, s (which estimates σ (sigma)), and summarize the data with a test statistic.

> The **test statistic** for the *t*-test for the population mean is:
>
> $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Recall that such a standardized test statistic represents how many standard deviations above or below μ_0 (mu_zero) our sample mean x-bar is.

Therefore our test statistic is a measure of how different our data are from what is claimed in the null hypothesis. This is an idea that we mentioned in the previous test as well.

Again we will rely on the **p-value to determine how unusual our data would be if the null hypothesis is true.**

As we mentioned, the test statistic in the *t*-test for a population mean does not follow a standard normal distribution. Rather, it follows another bell-shaped distribution called the *t*-distribution.

We will present the details of this distribution at the end for those interested but for now we will work on the process of the test.

Here are a few important facts.

- In statistical language we say that the **null distribution of our test statistic is the *t*-distribution with (n-1) degrees of freedom.** In other words, when Ho is true (i.e., when $\mu = \mu_0$ (mu = mu_zero)), our test statistic has a *t*-distribution with (n-1) d.f., and this is the distribution under which we find p-values.
- For a large sample size (n), the null distribution of the test statistic is approximately Z, so whether we use $t(n-1)$ or Z to calculate the p-values does not make a big difference. However, software will use the *t*-distribution regardless of the sample size and so will we.

Although we will not calculate p-values by hand for this test, we can still easily calculate the test statistic.

Try it yourself:

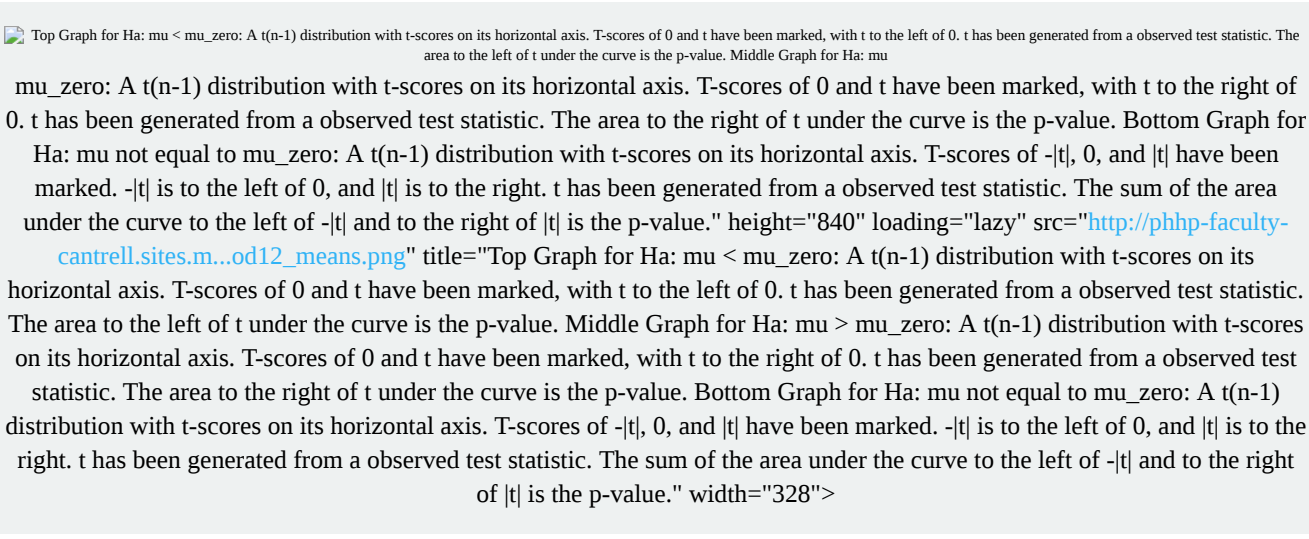
Learn by Doing: Calculate the Test Statistic for a Test for a Population Mean

From this point in this course and certainly in practice we will allow the software to calculate our test statistics and we will use the p-values provided to draw our conclusions.

Step 3: Find the p-value of the test by using the test statistic as follows

We will use software to obtain the p-value for this (and all future) tests but here are the images illustrating how the p-value is calculated in each of the three cases corresponding to the three choices for our alternative hypothesis.

Note that due to the symmetry of the t distribution, for a given value of the test statistic t, the p-value for the two-sided test is twice as large as the p-value of either of the one-sided tests. The same thing happens when p-values are calculated under the t distribution as when they are calculated under the Z distribution.



We will show some examples of p-values obtained from software in our examples. For now let's continue our summary of the steps.

Step 4: Conclusion

As usual, based on the p-value (and some significance level of choice) we assess the statistical significance of results, and draw our conclusions in context.

To review what we have said before:

If p-value ≤ 0.05 then **WE REJECT** Ho

- Conclusion: There **IS**enough evidence that *Ha is True*

If p-value > 0.05 then **WE FAIL TO REJECT** Ho

- Conclusion: There **IS NOT**enough evidence that *Ha is True*

Where instead of *Ha is True*, we write what this means in the words of the problem, in other words, in the context of the current scenario.

This step has essentially two sub-steps:

(i) Based on the p-value, **determine** whether or not the results are statistically **significant** (i.e., the data present enough evidence to reject Ho).

(ii) State your **conclusions** in the **context** of the problem.

We are now ready to look at two examples.

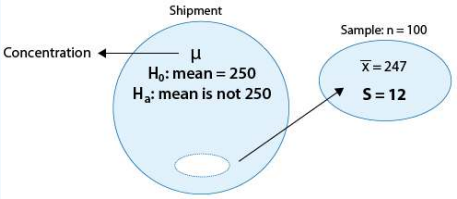
EXAMPLE:

A certain prescription medicine is supposed to contain an average of 250 parts per million (ppm) of a certain chemical. If the concentration is higher than this, the drug may cause harmful side effects; if it is lower, the drug may be ineffective.

The manufacturer runs a **check to see if the mean concentration in a large shipment conforms to the target level of 250 ppm or not.**

A simple random sample of 100 portions is tested, and the sample mean concentration is found to be 247 ppm with a sample standard deviation of 12 ppm.

Here is a figure that represents this example:



1. The hypotheses being tested are:

- Ho: $\mu = \mu_0$ (mu = mu_zero)
- Ha: $\mu \neq \mu_0$ (mu ≠ mu_zero)
- Where μ = population mean part per million of the chemical in the entire shipment

2. The conditions that allow us to use the t-test are met since:

- The **sample is random**
- The **sample size is large enough** for the Central Limit Theorem to apply and ensure the normality of x-bar. We do not need normality of the population in order to be able to conduct this test for the population mean. We are in the 2nd column in the table below.

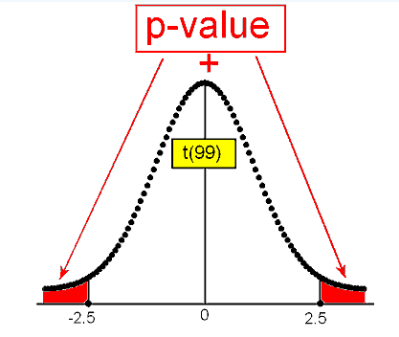
Conditions: t-test for a population mean	Small sample size	Large sample size
Variable varies normally in the population	✓	✓
Variable doesn't vary normally in the population	✗	✓

- The test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{247 - 250}{12/\sqrt{100}} = -2.5$$

- The data (represented by the sample mean) are 2.5 standard errors below the null value.

3. Finding the p-value.

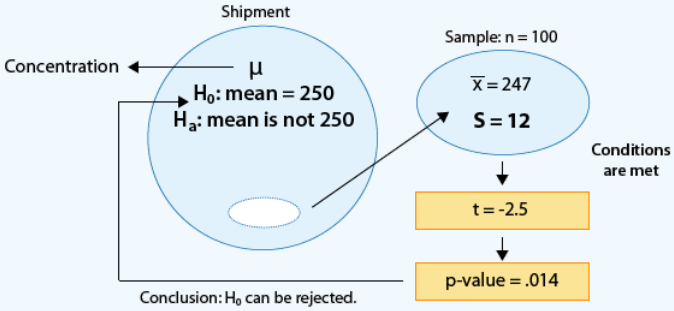


- To find the p-value we use statistical software, and we calculate a p-value of 0.014.

4. Conclusions:

- The p-value is small (.014) indicating that at the 5% significance level, the results are significant.
- We reject the null hypothesis.
- **OUR CONCLUSION IN CONTEXT:**
  - There **is enough evidence** to conclude that **the mean concentration in entire shipment is not the required 250 ppm.**
  - It is difficult to comment on the practical significance of this result without more understanding of the practical considerations of this problem.

Here is a summary:



Comments:

- The 95% confidence interval for μ (mu) can be used here in the same way as for proportions to conduct the two-sided test (checking whether the null value falls inside or outside the confidence interval) or following a *t*-test where Ho was rejected to get insight into the value of μ (mu).
- We find the **95% confidence interval to be (244.619, 249.381)**. Since 250 is not in the interval we know we would reject our null hypothesis that μ (mu) = 250. The confidence interval gives additional information. By accounting for estimation error, it estimates that the population mean is likely to be between 244.62 and 249.38. This is lower than the target concentration and that information might help determine the seriousness and appropriate course of action in this situation.

Caution

In most situations in practice we use TWO-SIDED HYPOTHESIS TESTS, followed by confidence intervals to gain more insight.

For completeness in covering one sample t-tests for a population mean, we still cover all three possible alternative hypotheses here HOWEVER, this will be the last test for which we will do so.

EXAMPLE:

A research study measured the pulse rates of 57 college men and found a mean pulse rate of 70 beats per minute with a standard deviation of 9.85 beats per minute.

Researchers want to know if the mean pulse rate for all college men is different from the current standard of 72 beats per minute.

1. **The hypotheses being tested are:**

- Ho: μ = 72
- Ha: μ ≠ 72
- Where μ = population mean heart rate among college men

2. **The conditions that allow us to use the *t*-test are met since:**

- The **sample is random.**
- The **sample size is large** (n = 57) so we do not need normality of the population in order to be able to conduct this test for the population mean. We are in the 2nd column in the table below.

Conditions: t-test for a population mean	Small sample size	Large sample size
Variable varies normally in the population	✓	✓
Variable doesn't vary normally in the population	✗	✓

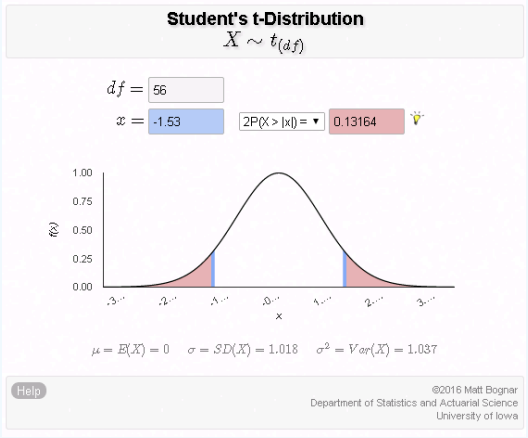
- The **test statistic** is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{70 - 72}{9.85/\sqrt{57}} = -1.53$$

- The data (represented by the sample mean) are 1.53 estimated standard errors below the null value.

3. Finding the p-value.

- Recall that in general the p-value is calculated under the null distribution of the test statistic, which, in the *t*-test case, is *t*(n-1). In our case, in which n = 57, the p-value is calculated under the *t*(56) distribution. Using statistical software, we find that the **p-value is 0.132**.
- Here is how we calculated the p-value. http://homepage.stat.uiowa.edu/~mbognar/applets/t.html.



4. Making conclusions.

- The p-value (0.132) is not small, indicating that the results are not significant.
- We fail to reject the null hypothesis.
- **OUR CONCLUSION IN CONTEXT:**
  - There is **not enough evidence** to conclude that the **mean pulse rate for all college men is different from the current standard of 72 beats per minute.**
  - The results from this sample do not appear to have any practical significance either with a mean pulse rate of 70, this is very similar to the hypothesized value, relative to the variation expected in pulse rates.

Now try a few yourself.

Learn by Doing: Hypothesis Testing for the Population Mean

From this point in this course and certainly in practice we will allow the software to calculate our test statistic and p-value and we will use the p-values provided to draw our conclusions.

That concludes our discussion of hypothesis tests in Unit 4A.

In the next unit we will continue to use both confidence intervals and hypothesis test to investigate the relationship between two variables in the cases we covered in Unit 1 on exploratory data analysis – we will look at Case CQ, Case CC, and Case QQ.

Before moving on, we will discuss the details about the *t*-distribution as a general object.

The t-Distribution

We have seen that variables can be visually modeled by many different sorts of shapes, and we call these shapes distributions. Several distributions arise so frequently that they have been given special names, and they have been studied mathematically.

So far in the course, the only one we've named, for continuous quantitative variables, is the normal distribution, but there are others. One of them is called the *t*-distribution.

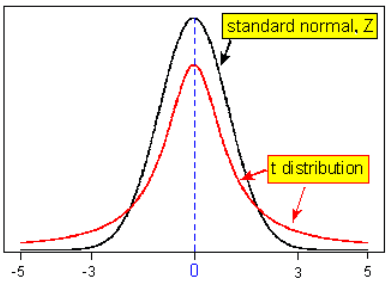
The *t*-distribution is another bell-shaped (unimodal and symmetric) distribution, like the normal distribution; and the center of the *t*-distribution is standardized at zero, like the center of the standard normal distribution.

Like all distributions that are used as probability models, the normal and the *t*-distribution are both scaled, so the total area under each of them is 1.

So how is the t-distribution fundamentally **different** from the normal distribution?

- The **spread**.

The following picture illustrates the fundamental difference between the normal distribution and the t-distribution:



Here we have an image which illustrates the fundamental difference between the normal distribution and the *t*-distribution:

You can see in the picture that the *t*-distribution has **slightly less area near the expected central value** than the normal distribution does, and you can see that the t distribution has correspondingly **more area in the "tails"** than the normal distribution does. (It's often said that the *t*-distribution has "fatter tails" or "heavier tails" than the normal distribution.)

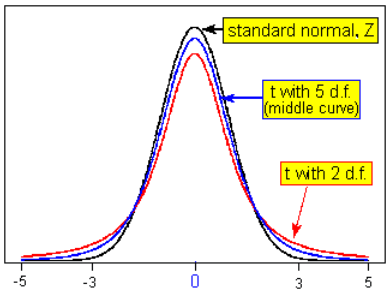
This reflects the fact that the *t*-distribution **has a larger spread** than the normal distribution. The same total area of 1 is spread out over a slightly wider range on the *t*-distribution, making it a bit lower near the center compared to the normal distribution, and giving the *t*-distribution slightly more probability in the 'tails' compared to the normal distribution.

Therefore, the *t*-distribution ends up being the appropriate model in certain cases where there is **more variability** than would be predicted by the normal distribution. One of these cases is stock values, which have more variability (or "volatility," to use the economic term) than would be predicted by the normal distribution.

There's actually an entire family of *t*-distributions. They all have similar formulas (but the math is beyond the scope of this introductory course in statistics), and they all have slightly "fatter tails" than the normal distribution. But some are closer to normal than others.

The *t*-distributions that have higher "degrees of freedom" are closer to normal (degrees of freedom is a mathematical concept that we won't study in this course, beyond merely mentioning it here). So, there's a *t*-distribution "with one degree of freedom," another *t*-distribution "with 2 degrees of freedom" which is slightly closer to normal, another *t*-distribution "with 3 degrees of freedom" which is a bit closer to normal than the previous ones, and so on.

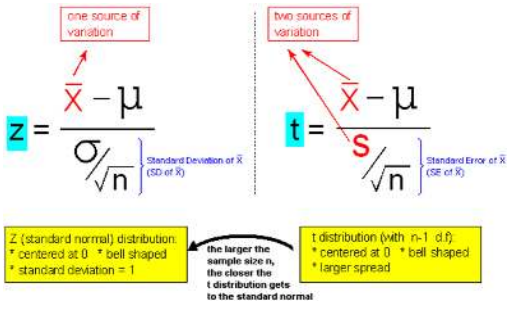
The following picture illustrates this idea with just a couple of *t*-distributions (note that "degrees of freedom" is abbreviated "d.f." on the picture):



The test statistic for our t-test for one population mean is a *t*-score which follows a *t*-distribution with (n – 1) degrees of freedom. Recall that each *t*-distribution is indexed according to "degrees of freedom." Notice that, in the context of a test for a mean, the degrees of freedom depend on the sample size in the study.

Remember that we said that higher degrees of freedom indicate that the *t*-distribution is closer to normal. So in the context of a test for the mean, the **larger the sample size**, the higher the degrees of freedom, and **the closer the *t*-distribution is to a normal z distribution**.

As a result, in the context of a test for a mean, the effect of the *t*-distribution is **most important** for a study with a **relatively small sample size**.



We are now done introducing the t-distribution. What are implications of all of this?

- The null distribution of our t-test statistic is the t-distribution with (n-1) d.f. In other words, when Ho is true (i.e., when $\mu = \mu_0$ (mu = mu_zero)), our test statistic has a t-distribution with (n-1) d.f., and this is the distribution under which we find p-values.
- For a large sample size (n), the null distribution of the test statistic is approximately Z, so whether we use t(n – 1) or Z to calculate the p-values does not make a big difference.

Wrap-Up (Inference for One Variable)

Video

Video: Summary Examples Unit 4A (34:51)

We've now completed the two main sections about inference for one variable. In these sections we introduced the three forms of inference:

- Point estimation—estimating an unknown parameter with a single value
- Interval estimation—estimating an unknown parameter with a confidence interval (an interval of plausible values for the parameter, which with some level of confidence we believe captures the true value of the parameter in it).
- Hypothesis testing — a four-step process in which we are assessing the statistical evidence provided by the data in favor or against some claim about the population.

Much like in the Exploratory Data Analysis section for one variable, we distinguished between the case when the variable of interest is categorical, and the case when it is quantitative.

- When the variable of interest is categorical, we are making an inference about the population proportion (p), which represents the proportion of the population that falls into one of the categories of the variable of interest.
- When the variable of interest is quantitative, the inference is about the population mean (μ, mu).

CHAPTER OVERVIEW

Unit 4B: Inference for Relationships

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.20: Classify a data analysis situation involving two variables according to the "role-type classification."

Learning Objectives

LO 4.35: For a data analysis situation involving two variables, choose the appropriate inferential method for examining the relationship between the variables and justify the choice.

Learning Objectives

LO 4.36: For a data analysis situation involving two variables, carry out the appropriate inferential method for examining relationships between the variables and draw the correct conclusions in context.

REVIEW: Unit 1 Role-Type Classification before continuing.

Video

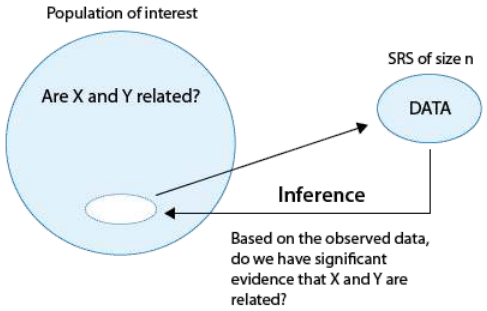
Video: Unit 4B: Inference for Relationships (5:15)

In the previous unit, we learned to perform inference for a **single** categorical or quantitative **variable** in the form of **point estimation**, **confidence intervals** or **hypothesis testing**.

The inference was actually

- about the **population proportion** (when the variable of interest was **categorical**) and
- about the **population mean** (when the variable of interest was **quantitative**).

Our next (and final) goal for this course is to perform **inference** about **relationships** between **two variables** in a population, based on an observed relationship between variables in a sample. Here is what the process looks like:



We are interested in studying whether a **relationship** exists **between** the **variables** X and Y **in a population of interest**. We choose a random sample and collect data on both variables from the subjects.

Our goal is to determine whether these data provide strong enough evidence for us to **generalize** the **observed relationship** in the **sample** and **conclude** (with some acceptable and agreed-upon level of uncertainty) that a **relationship** between X and Y **exists** in the entire **population**.

The primary form of inference that we will use in this unit is **hypothesis testing** but we will discuss **confidence intervals** both to estimate unknown parameters of interest involving two variables and as an alternative way of determining the conclusion to our hypothesis test.

> Conceptually, across all the inferential methods that we will learn, we'll test some form of:
>
> ***Ho: There is no relationship between X and Y***
>
> ***Ha: There is a relationship between X and Y***
>
> (We will also discuss point and interval estimation, but our discussion about these forms of inference will be framed around the test.)

Recall that when we discussed examining the relationship between two variables in the **Exploratory Data Analysis** unit, our discussion was framed around the **role-type classification**. This part of the course will be structured exactly in the same way.

In other words, we will look at hypothesis testing in the 3 sections corresponding to cases C→Q, C→C, and Q→Q in the table below.

		Response	
		Categorical	Quantitative
Explanatory	Categorical	**C→C**	**C→Q**
	Quantitative	**Q→C**	**Q→Q**

Recall that case Q→C is not specifically addressed in this course other than that we may investigate the association between these variables using the same methods as case C→Q.

It is also important to remember what we learned about lurking variables and causation.

- If our explanatory variable was part of a **well-designed experiment** then it may be **possible** for us to claim a **causal effect**
- But if it was based upon an **observational study**, we must be **cautious** to **imply only** a **relationship** or **association** between the two variables, **not** a direct **causal link** between the explanatory and response variable.

Unlike the previous part of the course on Inference for One Variable, where we discussed in some detail the theory behind the machinery of the test (such as the null distribution of the test statistic, under which the p-values are calculated), in the inferential procedures that we will introduce in Inference for Relationships, we will discuss much less of that kind of detail.

The principles are the same, but the details behind the null distribution of the test statistic (under which the p-value is calculated) become more complicated and require knowledge of theoretical results that are beyond the scope of this course.

Instead, **within each of the inferential methods we will focus on:**

- When the inferential method is appropriate for use.
- Under what conditions the procedure can safely be used.
- The conceptual idea behind the test (as it is usually captured by the test statistic).
- How to use software to carry out the procedure in order to get the p-value of the test.
- Interpreting the results in the context of the problem.
- Also, we will continue to introduce each test according to the four-step process of hypothesis testing.

Two-Sided Tests

From this point forward, we will generally focus on

- **TWO-SIDED tests** and
- **Supplement** with **confidence intervals** for the **effect of interest** to give further information

Using two-sided tests is **standard practice in clinical research** EVEN when there is a direction of interest for the research hypothesis, such as the desire to prove a new treatment is better than the current treatment.

Here are a few comments:

- Although fewer participants are required for one-sided tests, we are **unable to draw appropriate conclusions** if the study demonstrates the new treatment is worse. (See Defending the Rationale for the Two-Tailed Test in Clinical Research for a detailed discussion of this and other issues.)
- Using a one-sided test for the purpose of gaining statistical significance is **NOT A VALID APPROACH**. (See What are the differences between one-tailed and two-tailed tests? for more on this as well as a general overview of both types of tests.)

We are now ready to start with Case C→Q.

Topic hierarchy

Case C→C

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.35: For a data analysis situation involving two variables, choose the appropriate inferential method for examining the relationship between the variables and justify the choice.

Learning Objectives

LO 4.36: For a data analysis situation involving two variables, carry out the appropriate inferential method for examining relationships between the variables and draw the correct conclusions in context.

CO-5: Determine preferred methodological alternatives to commonly used statistical methods when assumptions are not met.

Review: Unit 1 Case C-C

Video

Video: Case C→C (47:09)

Related SAS Tutorials

- 6A – (3:07) Two-Way (Contingency) Tables – EDA
- 6B – (9:41) Two-Way (Contingency) Tables – Inference

Related SPSS Tutorials

- 6A – (7:57) Two-Way (Contingency) Tables – EDA
- 6B – (9:19) Two-Way (Contingency) Tables – Inference

Introduction

The last procedures we studied (two-sample t, paired t, ANOVA, and their non-parametric alternatives) all involve the relationship between a categorical explanatory variable and a quantitative response variable (case C→Q).In all of these procedures, the result is a comparison of the quantitative response variable (Y) among the groups defined by the categorical explanatory variable (X).The standard tests result in a comparison of the population means of Y within each group defined by X.

Next, we will consider inferences about the relationships between two categorical variables, corresponding to case C→C.

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C→C	✓C→Q
	Quantitative	Q→C	Q→Q

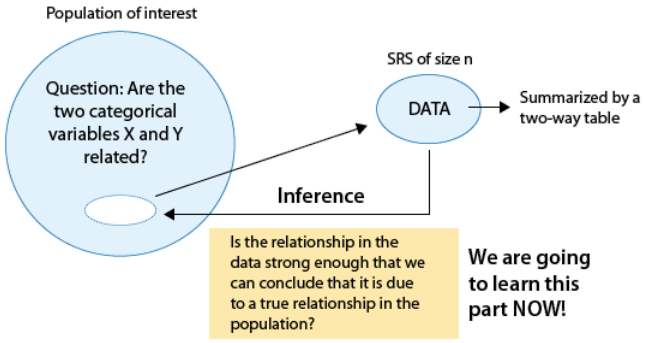
For case C→C, we will learn the following tests:

Independent Samples (Only Emphasis)	Dependent Samples (Not Discussed)
Standard Tests • **Continuity Corrected Chi-square Test for Independence (2×2 case)** • **Chi-square Test for Independence (RxC case)** Non-Parametric Test • Fisher's exact test	*Standard Test* • *McNemar's Test – 2×2 Case*

In the Exploratory Data Analysis unit of the course, we summarized the relationship between two categorical variables for a given data set (using a two-way table and conditional percents), without trying to generalize beyond the sample data.

Now we will perform statistical inference for two categorical variables, using the sample data to draw conclusions about whether or not we have evidence that the variables are related in the larger population from which the sample was drawn.

In other words, we would like to **assess whether the relationship between X and Y that we observed in the data is due to a real relationship between X and Y in the population, or if it is something that could have happened just by chance due to sampling variability.**



Before moving into the statistical tests, let's look at a few (fake) examples.

RxC Tables

Suppose our explanatory variable X has r levels and our response variable Y has c levels. We usually arrange our table with the explanatory variable in the rows and the response variable in the columns.

EXAMPLE: RxC Table

Suppose we have the following partial (fake) data summarized in a two-way table using X = BMI category (r = 4 levels) and Y = Diabetes Status (c = 3 levels).

	No Diabetes	Pre-Diabetes	Diabetes	Total
Underweight				100
Normal				400
Overweight				300
Obese				200
Total	700	200	100	1000

From our study of probability we can determine:

- P(No Diabetes) = 700/1000 = 0.7
- P(Pre-Diabetes) = 200/1000 = 0.20
- P(Diabetes) = 100/1000 = 0.10

In the test we are going to use, our **null hypothesis** will be:

Ho: There is no relationship between X and Y.

Which in this case would be:

Ho: There is no relationship between BMI category (X) and diabetes status (Y).

If there were no relationship between X and Y, this would imply that the distribution of diabetes status is the same for each BMI category.

In this case (C→C), the distribution of diabetes status consists of the probability of each diabetes status group and the null hypothesis becomes:

Ho: BMI category (X) and diabetes status (Y) are **INDEPENDENT**.

Since the probability of "No Diabetes" is 0.7 in the entire dataset, if there were no differences in the distribution of diabetes status between BMI categories, we would obtain the same proportion in each row. Using the row totals we can find the **EXPECTED** counts as follows.

Notice the formula used below is simply the formula for the mean or expected value of a binomial random variable with n "trials" and probability of "success" p which was $\mu = E(X) = np$ where X = number of successes for a sample of size n.

	No Diabetes	Pre-Diabetes	Diabetes	Total
Underweight	100(0.7) = 70			100
Normal	400(0.7) = 280			400
Overweight	300(0.7) = 210			300
Obese	200(0.7) = 140			200
Total	700	200	100	1000

Notice that these do indeed add to 700.

Similarly we can determine the **EXPECTED** counts for the remaining two columns since 20% of our sample were classified as having pre-diabetes and 10% were classified as having diabetes.

	No Diabetes	Pre-Diabetes	Diabetes	Total
Underweight	70	100(0.2) = 20	100(0.1) = 10	100
Normal	280	400(0.2) = 80	400(0.1) = 40	400
Overweight	210	300(0.2) = 60	300(0.1) = 30	300
Obese	140	200(0.2) = 40	200(0.1) = 20	200
Total	700	200	100	1000

What we have created, using only the row totals, column totals, and column percents, is a table of what we would expect to happen if the null hypothesis of no relationship between X and Y were true. Here is the final result.

	No Diabetes	Pre-Diabetes	Diabetes	Total
Underweight	**70**	**20**	**10**	100
Normal	**280**	**80**	**40**	400
Overweight	**210**	**60**	**30**	300
Obese	**140**	**40**	**20**	200
Total	700	200	100	1000

Suppose we gather data and find the following (expected counts are in parentheses for easy comparison):

	No Diabetes	Pre-Diabetes	Diabetes	Total
Underweight	**65** (*70*)	**22** (*20*)	**13** (*10*)	100
Normal	**285** (*280*)	**78** (*80*)	**37** (*40*)	400
Overweight	**216** (*210*)	**53** (*60*)	**31** (*30*)	300
Obese	**134** (*140*)	**47** (*40*)	**19** (*20*)	200
Total	700	200	100	1000

If we compare our counts to the expected counts they are fairly close. This data would not give much evidence of a difference in the distribution of diabetes status among the levels of BMI categories. In other words, this data would not give much evidence of a relationship (or association) between BMI categories and diabetes status.

The standard test we will learn in case C→C is based upon comparing the **OBSERVED** cell counts (our data) to the **EXPECTED** cell counts (using the method discussed above).

We want you to see how the expected cell counts are created so that you will understand what kind of evidence is being used to reject the null hypothesis in case C→C.

Suppose instead that we gather data and we obtain the following counts (expected counts are in parentheses and row percentages are provided):

	No Diabetes	Pre-Diabetes	Diabetes	Total
Underweight	**90** (*70*) 90%	**7** (*20*) 7%	**3** (*10*) 3%	100
Normal	**340** (*280*) 85%	**40** (*80*) 10%	**20** (*40*) 5%	400
Overweight	**180** (*210*) 60%	**90** (*60*) 30%	**30** (*30*) 10%	300
Obese	**90** (*140*) 45%	**63** (*40*) 31.5%	**47** (*20*) 23.5%	200
Total	700	200	100	1000

In this case, most of the differences are drastic and there seems to be clear evidence that the distribution of diabetes status is not the same among the four BMI categories.

Although this data is entirely fabricated, it illustrates the kind of evidence we need to reject the null hypothesis in case C→C.

2×2 Tables

One special case occurs when we have two categorical variables where both of these variables have two levels. Two-level categorical variables are often called **binary** variables or **dichotomous** variables and when possible are usually coded as 1 for "Yes" or "Success" and 0 for "No" or "Failure."

Here is another (fake) example.

EXAMPLE: 2x2 Table

Suppose we have the following partial (fake) data summarized in a two-way table using X = treatment and Y = significant improvement in symptoms.

	No Improvement	Improvement	Total

Control			100
Treatment			100
Total	120	80	200

From our study of probability we can determine:

- P(No Improvement) = 120/200 = 0.6
- P(Improvement) = 80/200 = 0.4

Since the probability of "No Improvement" is 0.6 in the entire dataset and the probability for "Improvement" is 0.4, if there was no difference we would obtain the same proportion in each row. Using the row totals we can find the EXPECTED counts as follows.

	No Improvement	Improvement	Total
Control	**100(0.6) = 60**	**100(0.4) = 40**	100
Treatment	**100(0.6) = 60**	**100(0.4) = 40**	100
Total	120	80	200

Suppose we obtain the following data:

	No Improvement	Improvement	Total
Control	**80**	**20**	100
Treatment	**40**	**60**	100
Total	120	80	200

In this example we are interested in the probability of improvement and the above data seem to indicate the treatment provides a greater chance for improvement than the control.

We use this example to mention two ways of comparing probability (sometimes "risk") in 2×2 tables. Many of you may remember these topics from Epidemiology or may see these topics again in Epidemiology courses in the future!

Risk Difference:

For this data, a larger proportion of subjects in the treatment group showed improvement compared to the control group. In fact, the estimated probability of improvement is 0.4 higher for the treatment group than the control group.

This value (0.4) is called a **risk-difference** and is one common measure in 2×2 tables. Estimates and confidence intervals can be obtained.

For a fixed sample size, the larger this difference, the more evidence against our null hypothesis (no relationship between X and Y).

The population risk-difference is often denoted $p_1 - p_2$, and is the difference between two population proportions. We estimate these proportions in the same manner as Unit 1, once for each sample.

For the current example, we obtain

$$\hat{p}_1 = \hat{p}_{\text{TRT}} = \frac{60}{100} = 0.60$$

and

$$\hat{p}_2 = \hat{p}_{\text{Control}} = \frac{20}{100} = 0.20$$

from which we find the risk difference

$$\hat{p}_{\text{TRT}} - \hat{p}_{\text{Control}} = 0.60 - 0.20 = 0.40$$

Odds Ratio:

Another common measure in 2×2 tables is the odds ratio, which is defined as the odds of the event occurring in one group divided by the odds of the event occurring in another group.

In this case, the odds of improvement in the treatment group is

$$\text{ODDS}_{\text{TRT}} = \frac{P(\text{ Improvement } \mid \text{TRT})}{P(\text{ No Improvement } \mid \text{TRT})} = \frac{0.6}{0.4} = 1.5$$

and the odds of improvement in the control group is

$$\text{ODDS}_{\text{Control}} = \frac{P(\text{ Improvement } \mid \text{ Control })}{P(\text{ No Improvement } \mid \text{ Control })} = \frac{0.2}{0.8} = 0.25$$

so the odds ratio to compare the treatment group to the control group is

$$\text{Odds Ratio } = \frac{\text{ODDS }_{\text{TRT}}}{\text{ODDS}_{\text{Control}}} = \frac{1.5}{0.25} = 6$$

This value means that the odds of improvement are 6 times higher in the treatment group than in the control group.

Properties of Odds Ratios:

- The odds ratio is always larger than 0.
- An odds ratio of 1 implies the odds are equal in the two groups.
- Values much larger than 1 indicate the event is more likely in the treatment group (numerator group) than the control group (denominator group). This would give evidence that our null hypothesis is false.
- Values much smaller than 1 (closer to zero) would indicate the event is much less likely in the treatment group than the control group. This would also give evidence that our null hypothesis is false.
    - **Notice:** if we compared control to treatment (instead of treatment to control) we would obtain an odds ratio of 1/6 which would say that the odds of improvement in the control group is 1/6 the odds of improvement in the treatment group which leads us to exactly the same conclusion, worded in an opposite manner.

Chi-square Test for Independence

Learning Objectives

LO 4.43: In a given context, determine the appropriate standard method for examining the relationship between two categorical variables. Given the appropriate software output choose the correct p-value and provide the correct conclusions in context.

Learning Objectives

LO 4.44: In a given context, set up the appropriate null and alternative hypotheses for examining the relationship between two categorical variables.

Step 1: State the hypotheses The hypotheses are:

Ho: There is no relationship between the two categorical variables. (They are independent.)

Ha: There is a relationship between the two categorical variables. (They are not independent.)

Note: for 2×2 tables, these hypotheses can be formulated the same as for population means except using population proportions. This can be done for RxC tables as well but is not common as it requires more notation to compare multiple group proportions.

- **Ho:** $p_1 - p_2 = 0$ (which is the same as $p_1 = p_2$)
- **Ha:** $p_1 - p_2 \neq 0$ (which is the same as $p_1 \neq p_2$) **(two-sided)**

Step 2: Obtain data, check conditions, and summarize data

(i) The sample should be random with independent observations (all observations are independent of all other observations).

(ii) In general, the larger the sample, the more precise and reliable the test results are. There are different versions of what the conditions are that will ensure reliable use of the test, all of which involve the expected counts. One version of the conditions says that all expected counts need to be greater than 1, and at least 80% of expected counts need to be greater than 5. A more conservative version requires that all expected counts are larger than 5. Some software packages will provide a warning if the sample size is "too small."

Test Statistic of the Chi-square Test for Independence:

The single number that summarizes the overall difference between observed and expected counts is the chi-square statistic, which tells us in a standardized way how far what we observed (data) is from what would be expected if Ho were true.

Here it is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{ observed count } - \text{ expected count })^2}{\text{expected count}}$$

Step 3: Find the p-value of the test by using the test statistic as follows We will rely on software to obtain this value for us. We can also request the expected counts using software.

The p-values are calculated using a chi-square distribution with (r-1)(c-1) degrees of freedom (where r = number of levels of the row variable and c = number of levels of the column variable). We will rely on software to obtain the p-value for this test.

IMPORTANT NOTE

- **Use Continuity Correction for 2×2 Tables:** For 2×2 tables, a continuity correction is used to improve the approximation of the p-value. This value will only be calculated by the software for 2×2 tables where both variables are binary – have only two levels.

Step 4: Conclusion

As usual, we use the magnitude of the p-value to draw our conclusions. A small p-value indicates that the evidence provided by the data is strong enough to reject Ho and conclude (beyond a reasonable doubt) that the two variables are related. In particular, if a significance level of 0.05 is used, we will reject Ho if the p-value is less than 0.05.

Non-Parametric Alternative: Fisher's Exact Test

Learning Objectives

LO 5.1: For a data analysis situation involving two variables, determine the appropriate alternative (non-parametric) method when assumptions of our standard methods are not met.

We will look at one non-parametric test in case C→C. Fisher's exact test is an exact method of obtaining a p-value for the hypotheses tested in a standard chi-square test for independence. This test is often used when the sample size requirement of the chi-square test is not satisfied and can be used for 2×2 and RxC tables.

Step 1: State the hypotheses The hypotheses are:

Ho: There is no relationship between the two categorical variables. (They are independent.)

Ha: There is a relationship between the two categorical variables. (They are not independent, they are dependent.)

Step 2: Obtain data, check conditions, and summarize data

The sample should be random with independent observations (all observations are independent of all other observations).

Step 3: Find the p-value of the test by using the test statistic as follows

The p-values are calculated using a distribution specific to this test. We will rely on software to obtain the p-value for this test. The p-value measures the chance of obtaining a table as or more extreme (against the null hypothesis) than our table.

Step 4: Conclusion

As usual, we use the magnitude of the p-value to draw our conclusions. A small p-value indicates that the evidence provided by the data is strong enough to reject Ho and conclude (beyond a reasonable doubt) that the two variables are related. In particular, if a significance level of 0.05 is used, we will reject Ho if the p-value is less than 0.05.

Now let's look at a some examples with real data.

EXAMPLE: Risk Factor for Low Birth Weight

Low birth weight is an outcome of concern due to the fact that infant mortality rates and birth defect rates are very high for babies with low birth weight. A woman's behavior during pregnancy (including diet, smoking habits, and obtaining prenatal care) can greatly alter her chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

In this example, we will use a 1986 study (Hosmer and Lemeshow (2000), Applied Logistic Regression: Second Edition) in which data were collected from 189 women (of whom 59 had low birth weight infants) at the Baystate Medical Center in Springfield, MA. The goal of the study was to identify risk factors associated with giving birth to a low birth weight baby.

Data: SPSS format, SAS format, Excel format

Response Variable:

- LOW – Low birth weight
  - 0=No (birth weight >= 2500 g)
  - 1=Yes (birth weight < 2500 g)

Possible Explanatory Variables (variables we will use in this example are in bold):

- **RACE** – Race of mother (1=White, 2=Black, 3=Other)
- **SMOKE** – Smoking status during pregnancy (0=No, 1=Yes)
- **PTL** – History of premature labor (0=None, 1=One, etc.)
- **HT** – History of hypertension (0=No, 1=Yes)
- **UI** – Presence of uterine irritability (0=No, 1=Yes)
- FTV – Number of physician visits during the first trimester
- BWT – The actual birth weight (in grams)
- AGE – Age of mother (in years)
- LWT – Weight of mother at the last menstrual period (in pounds)

Results:

Step 1: State the hypotheses

The hypotheses are:

> **Ho**: There is no relationship between the categorical explanatory variable and presence of low birth weight. (They are independent.)
>
> **Ha**: There is a relationship between the categorical explanatory variable and presence of low birth weight.(They are not independent, they are dependent.)

Steps 2 & 3: Obtain data, check conditions, summarize data, and **find the p-value**

Explanatory Variable	Which Test is Appropriate?	P-value	Decision
RACE	Min. Expected Count = **8.12** **3×2 table** Use Pearson Chi-square (since RxC)	**0.0819** (Chi-square – SAS) **0.082** (Chi-square – SPSS)	Fail to Reject Ho

SMOKE	Min. Expected Count = **23.1** **2×2 table** Use Continuity Correction (since 2×2)	**0.040** (Continuity Correction – SPSS) **0.0396** (Continuity Adj – SAS)	Reject Ho
PTL	Min. Expected Count = **0.31** **4×2 table** Fisher's Exact test is more appropriate	**3.106 E-04 = 0.0003106** (Fisher's – SAS) **0.000** (Fisher's – SPSS) 0.0008 (Chi-square – SAS) 0.001 (Chi-square – SPSS)	Reject Ho
HT	Min. Expected Count = **3.75** **2×2 table** Fisher's Exact test may be more appropriate	**0.0516** (Fisher's – SAS) **0.052** (Fisher's – SPSS)	Fail to Reject Ho (Barely)
UI	Min. Expected Count = **8.74** **2×2 table** Use Continuity Correction	**0.0355** (Continuity Adj. – SAS) **0.035** (Continuity Correction – SPSS)	Reject Ho

Step 4: Conclusion

When considered individually, presence of uterine irritability, history of premature labor, and smoking during pregnancy are all significantly associated (p-value < 0.05) with the presence/absence of a low birth weight infant whereas history of hypertension and race were only marginally significant ($0.05 \leq$ p-value < 0.10).

Practical Significance:

Explanatory Variable	Comparison of Conditional Percentages of Low Birth Weight
RACE	Race = White: 23.96% Race = Black: 42.31% Race = Other: 37.31%
SMOKE	Smoke = No: 25.22% Smoke = Yes: 40.54%
PTL	History of Premature Labor = 0: 25.79% History of Premature Labor = 1: 66.67% History of Premature Labor = 2: 40.00% (Note small sample size of 5 for this row) History of Premature Labor = 3: 0.00% (Note small sample size of 1 for this row)
HT	Hypertension = No: 29.38% Hypertension = Yes: 58.33% (Note small sample size of 12 for this row)
UI	Presence of uterine irritability = No: 27.95% Presence of uterine irritability = Yes: 50.00%

- Despite our failing to reject the null in two of the five tests, all of these results seem to have some practical significance although the small sample sizes for some portions of the results may be producing misleading information and likely would require further study to confirm the results seen here.

SPSS Output for tests

SAS Output, SAS Code

EXAMPLE: 2x2 Table - Revisting "Looks vs. Personality" with Binary Categorized Response

If, instead of simply analyzing the “looks vs. personality” rating scale, we categorized the responses into groups then we would be in case C→C instead of case C→Q (see previous example in Case C-Q for Two Independent Samples).

Recall the rating score was from 1 to 25 with 1 = personality most important (looks not important at all) and 25 = looks most important (personality not important at all). A score of 13 would be equally important and scores around 13 should indicate looks and personality are nearly equal in importance.

For our purposes we will use a rating of 16 or larger to indicate that looks were indeed more important than personality (by enough to matter).

Data: SPSS format, SAS format

Response Variable:

- Looks – “Looks were (much) more important?”
  - 0=No (Less than 16 on the looks vs. personality rating)
  - 1=Yes (16 or higher on the looks vs. personality rating)

Results:

Step 1: State the hypotheses

The hypotheses are:

> **Ho:** The proportion of college students who find looks more important than personality **is the same** for males and females. (The two variables are independent)
>
> **Ha:** The proportion of college students who find looks more important than personality **is different** for males and females. (The two variables are dependent)

Steps 2 & 3: Obtain data, check conditions, summarize data, and f**ind the p-value**

The minimum expected cell count is 13.38. This is a 2×2 table so we will use the continuity corrected chi-square statistic.

The p-value is found to be 0.001 (SPSS) or 0.0007 (SAS).

Step 4: Conclusion

There is a significant association between gender and whether or not the individual rated looks more important than personality.

Among males, 27.1% rated looks higher than personality while among females this value was only 9.3%.

For fun: The odds ratio here is

$$\text{Odds Ratio} = \frac{0.271/(1-0.271)}{0.093/(1-0.093)} = \frac{0.37174}{0.10254} = 3.63$$

which means, based upon our data, we estimate that the odds of rating looks more important than personality is 3.6 times higher among males than among females.

Practical Significance:

It seems clear that the difference between 27.1% and 9.3% is practically significant as well as statistically significant. This difference is large and likely represents a meaningful difference in the views of males and females regarding the importance of looks compared to personality.

SPSS Output

SAS Output, SAS Code

Case C→Q

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

REVIEW: Unit 1 Case C-Q

Video

Video: Case C→Q (5:23)

Introduction

Recall the role-type classification table framing our discussion on inference about the relationship between two variables.

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C→C	**C→Q**
	Quantitative	Q→C	Q→Q

We start with case C→Q, where the explanatory variable is categorical and the response variable is quantitative.

Recall that in the Exploratory Data Analysis unit, examining the relationship between X and Y in this situation amounts, in practice, to:

- **Comparing the distributions of the (quantitative) response Y for each value (category) of the explanatory X.**

To do that, we used

- **side-by-side boxplots** (each representing the distribution of Y in one of the groups defined by X),
- and supplemented the display with the corresponding **descriptive statistics**.

We will need to add one layer of difficulty here with the possibility that we may have **paired** or **matched samples** as opposed to **independent samples** or **groups**. Note that all of the examples we discussed in Case CQ in Unit 1 consisted of independent samples.

First we will review the general scenario.

Comparing Means between Groups

To understand the logic, we'll start with an example and then generalize.

EXAMPLE: GPA and Year in College

Suppose that our variable of interest is the GPA of college students in the United States. From Unint 4A, we know that since GPA is **quantitative**, we will conduct inference on μ, the (**population**) **mean GPA** among all U.S. college students.

Since this section is about relationships, let's assume that what we are really interested in is not simply GPA, but the relationship between:

- **X : year in college** (1 = freshmen, 2 = sophomore, 3 = junior, 4 = senior) and
- **Y : GPA**

In other words, we want to explore whether **GPA** is **related** to **year in college**.

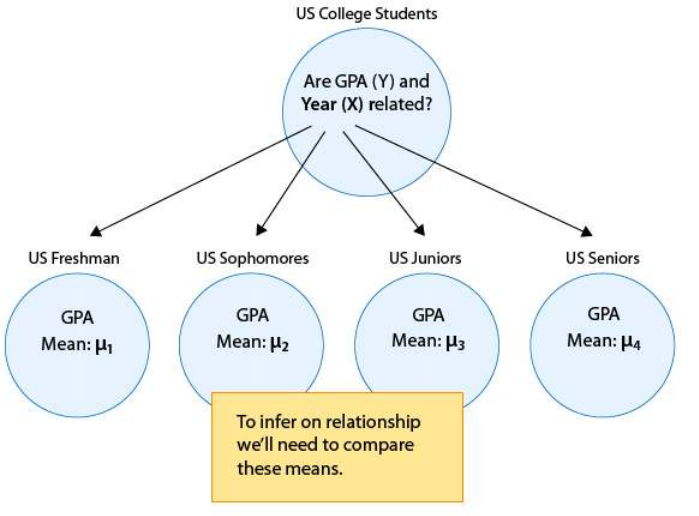
The way to think about this is that the population of U.S. college students is now broken into **4 sub-populations:** freshmen, sophomores, juniors and seniors. Within each of these four groups, we are interested in the GPA.

The inference must therefore involve the 4 sub-population means:

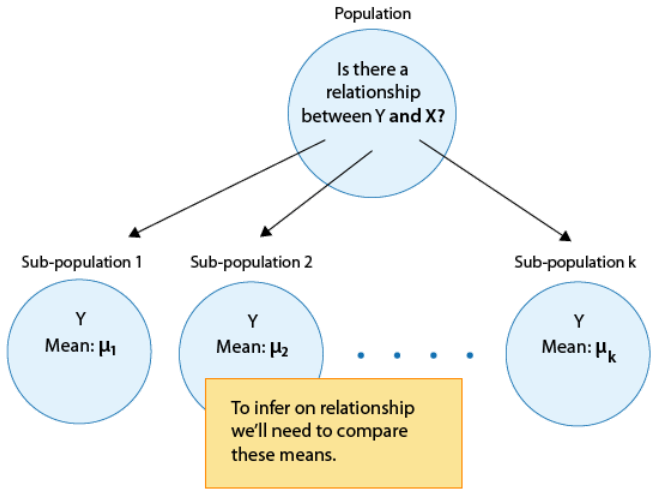
- μ_1 : mean GPA among **freshmen** in the United States.
- μ_2 : mean GPA among **sophomores** in the United States
- μ_3 : mean GPA among **juniors** in the United States
- μ_4 : mean GPA among **seniors** in the United States

It makes sense that the inference about the relationship between year and GPA has to be based on some kind of comparison of these four means.

If we infer that these four means are not all equal (i.e., that there are some differences in GPA across years in college) then that's equivalent to saying GPA is related to year in college. Let's summarize this example with a figure:



In general, making inferences about the relationship between X and Y in Case C→Q boils down to comparing the means of Y in the sub-populations, which are created by the categories defined by X (say k categories). The following figure summarizes this:



We will split this into two different scenarios (**k = 2** and **k > 2**), where k is the number of categories defined by X.

For example:

- If we are interested in whether GPA (Y) is related to **gender** (**X**), this is a scenario where **k = 2** (since gender has only two categories: M, F), and the inference will boil down to comparing the mean GPA in the sub-population of males to that in the

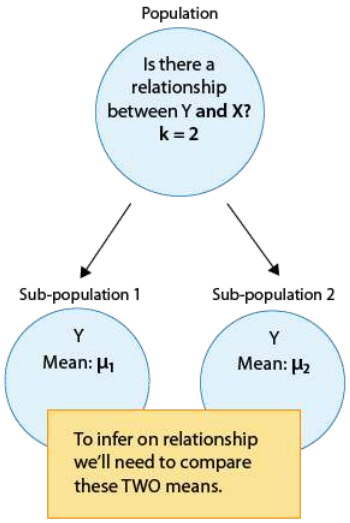
sub-population of females.

- On the other hand, in the example we looked at earlier, the relationship between GPA (Y) and **year in college** (**X**) is a scenario where **k > 2** or more specifically, k = 4 (since year has four categories).

Caution

In terms of inference, these two situations (k = 2 and k > 2) will be treated differently!

Scenario with k = 2



Scenario with k > 2

The entire population is represented by a large circle, for which we wonder if there is a relationship between Y and X. k 2. This large population is broken up into k sub-populations, each with its own mean μ. To infer on relationship between Y and X, we'll need to compare these k means." height="459" loading="lazy" src="http://phhp-faculty-cantrell.sites.m...7/image013.gif" title="The entire population is represented by a large circle, for which we wonder if there is a relationship between Y and X. k > 2. This large population is broken up into k sub-populations, each with its own mean μ. To infer on relationship between Y and X, we'll need to compare these k means." width="565">

Dependent vs. Independent Samples (k = 2)

Learning Objectives

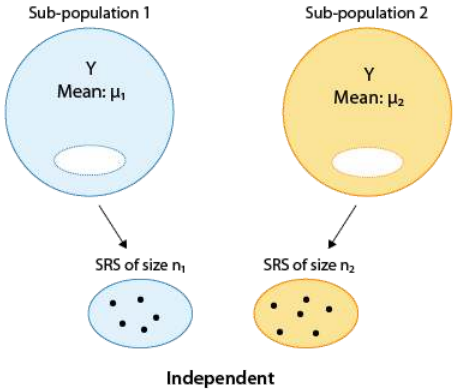
LO 4.37: Identify and distinquish between independent and dependent samples.

Furthermore, within the scenario of **comparing two means** (i.e., examining the relationship between X and Y, when X has only two categories, k = 2) we will distinguish between two scenarios.

Here, the distinction is somewhat subtle, and has to do with how the samples from each of the two sub-populations we're comparing are chosen. In other words, it depends upon **what type of study design** will be implemented.

We have learned that many experiments, as well as observational studies, make a comparison between two groups (sub-populations) defined by the categories of the explanatory variable (X), in order to see if the response (Y) differs.

In some situations, one group (sub-population 1) is defined by one category of X, and **another independent group** (sub-population 2) is defined by the other category of X. Independent samples are then taken from each group for comparison.



EXAMPLE:

Suppose we are conducting a clinical trial. Participants are randomized into two independent subpopulations:

- those who are given a drug and
- those who are given a placebo.

Each individual appears in only one of these two groups and individuals are not matched or paired in any way. Thus the two samples or groups are **independent**. We can say those given the drug are **independent** from those given the placebo.

Recall: By randomly assigning individuals to the treatment we control for both known and unknown lurking variables.

EXAMPLE:

Suppose the Highway Patrol wants to study the reaction times of drivers with a blood alcohol content of half the legal limit in their state.

An observational study was designed which would also serve as publicity on the topic of drinking and driving. At a large event where enough alcohol would be consumed to obtain plenty of potential study participants, officers set up an obstacle course and provided the vehicles. (Other considerations were also implemented to keep the car and track conditions consistent for each participant.)

Volunteers were recruited from those in attendance and given a breathalyzer test to determine their blood alcohol content. Two types of volunteers were chosen to participate:

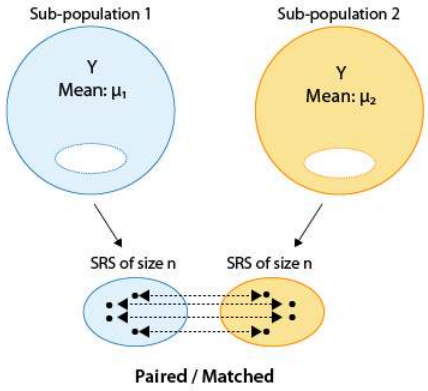
- Those with a blood alcohol content of zero – as measured by the breathalyzer – of which 10 were chosen to drive the course.
- Those with a blood alcohol content within a small range of half the legal limit (in Florida this would be around 0.04%) – of which 9 were chosen.

Here also, we have two **independent** groups – even if originally they were taken from the same sample of volunteers – each individual appears in only one of the two groups, the comparison of the reaction times is a comparison **between two independent groups.**

However, in this study, there **was NO random assignment** to the treatment and so we would need to be much more concerned about the possibility of lurking variables in this study compared to one in which individuals were randomized into one of these two groups.

We will see it may be more appropriate in some studies to use the same individual as a subject in BOTH treatments – this will result in **dependent samples**.

When a matched pairs sample design is used, each observation in one sample is **matched/paired/linked** with an observation in the other sample. These are sometimes called "**dependent samples**."



Matching could be by person (if the same person is measured twice), or could actually be a pair of individuals who belong together in a relevant way (husband and wife, siblings).

In this design, then, the **same individual** or a **matched pair** of individuals is **used** to **make two measurements** of the **response –** one for each of the **two levels** of the **categorical explanatory variable**.

Advantages of a paired sample approach include:

- Reduced measurement error since the variance within subjects is typically smaller than that between subjects
- Requires smaller number of subjects to achieve the same power than independent sample methods.

Disadvantages of a paired sample approach include:

- An order effect based upon which treatment individuals received first.
- A carryover effect such as a drug remaining in the system.

- Testing effect such as particpants learning the obstacle course in the first run improving their performance in the 2nd.

EXAMPLE:

Suppose we are conducting a study on a pain blocker which can be applied to the skin and are comparing two different dosage levels of the solution which in this study will be applied to the forearm.

For each participant both solutions are applied with the following protocol:

- Which drug is applied to which arm is random.
- Patients and clinical staff are blind to the two treatment applications.
- Pain tolerance is measured on both arms using the same standard test with the order of testing randomized.

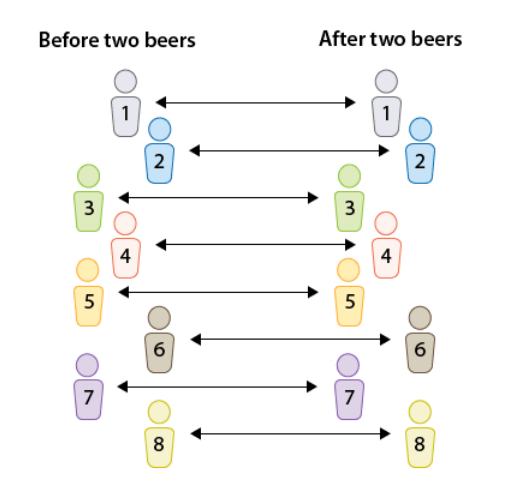
Here we have dependent samples since the same patient appears in both dosage groups.

Again, randomization is employed to help minimize other issues related to study design such as an order or testing effect.

EXAMPLE:

Suppose the department of motor vehicles wants to check whether drivers are impaired after drinking two beers.

The reaction times (measured in seconds) in an obstacle course are measured for 8 randomly selected drivers **before and then after** the consumption of two beers.



We have a matched-pairs design, since each individual was measured twice, once before and once after.

In matched pairs, the comparison between the reaction times is done **for each individual.**

Comment:

- Note that in the first figure, where the samples are independent, the sample sizes of the two independent samples need not be the same.
- On the other hand, it is obvious from the design that in the matched pairs the sample sizes of the two samples must be the same (and thus we used n for both).
- Dependent samples can occur in many other settings but for now we focus on the case of investigating the relationship between a two-level categorical explanatory variable and a quantitative response variable.

Let's Summarize:

We will begin our discussion of Inference for Relationships with Case C-Q, where the explanatory variable (X) is categorical and the response variable (Y) is quantitative. We discussed that inference in this case amounts to comparing population means.

- We distinguish between scenarios where the explanatory variable (X) has only two categories and scenarios wheret he explanatory variable (X) has MORE than two categories.
- When comparing two means, we make the futher distinction between situations where we have independent samples and those where we have matched pairs.
- For comparing more than two means in this course, we will focus only on the situation where we have independent samples. In studies with more than two groups on dependent samples, it is good to know that a common method used is repeated measures but we will not cover it here.
- We will first discuss comparing two population means starting with matched pairs (dependent samples) then independent samples and conclude with comparing more than two population means in the case of independent samples.

Now test your skills at identifying the three scenarios in Case C-Q.

> **Did I Get This?:** Scenarios in Case C-Q
> (Non-Interactive Version – Spoiler Alert)

Looking Ahead – Methods in Case C-Q

- Methods in **BOLD** will be our main focus in this unit.

Here is a summary of the tests we will learn for the scenario where k = 2.

Independent Samples (More Emphasis)	Dependent Samples (Less Emphasis)
Standard Tests • **Two Sample T-Test Assuming Equal Variances** • **Two Sample T-Test Assuming Unequal Variances** Non-Parametric Test • Mann-Whitney U (or Wilcoxon Rank-Sum) Test	**Standard Test** • **Paired T-Test** Non-Parametric Tests • Sign Test • Wilcoxon Signed-Rank Test

Here is a summary of the tests we will learn for the scenario where k > 2.

Independent Samples (Only Emphasis)	Dependent Samples (Not Discussed)
Standard Tests • **One-way ANOVA (Analysis of Variance)** Non-Parametric Test • Kruskal–Wallis One-way ANOVA	*Standard Test* • *Repeated Measures ANOVA (or similar)*

Paired Samples

Caution

As we mentioned at the end of the Introduction to Unit 4B, **we will focus only on two-sided tests** for the remainder of this course. One-sided tests are often possible but rarely used in clinical research.

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.35: For a data analysis situation involving two variables, choose the appropriate inferential method for examining the relationship between the variables and justify the choice.

Learning Objectives

LO 4.36: For a data analysis situation involving two variables, carry out the appropriate inferential method for examining relationships between the variables and draw the correct conclusions in context.

CO-5: Determine preferred methodological alternatives to commonly used statistical methods when assumptions are not met.

Video

Video: Paired Samples (27:19)

Related SAS Tutorials

- 8B (2:55) EDA of Differences
- 8C (5:20) Paired T-Test and Non Parametric Tests

Related SPSS Tutorials

- 8B (2:00) EDA of Differences
- 8C (3:11) Paired T-Test
- 8D (3:32) Non Parametric (Paired)

Introduction – Matched Pairs (Paired t-test)

Learning Objectives

LO 4.37: Identify and distinguish between independent and dependent samples.

Learning Objectives

LO 4.38: In a given context, determine the appropriate standard method for comparing groups and provide the correct conclusions given the appropriate software output.

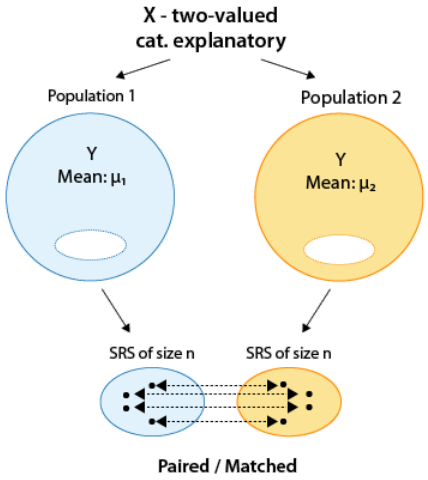
Learning Objectives

LO 4.39: In a given context, set up the appropriate null and alternative hypotheses for comparing groups.

We are in **Case CQ of inference about relationships**, where the **explanatory variable is categorical** and the **response variable is quantitative**.

As we mentioned in the summary of the introduction to Case C→Q, the first case that we will deal with is that involving **matched pairs**. In this case:

- The samples are paired or matched. Every observation in one sample is **linked** with an observation in the other sample.
- In other words, the samples are **dependent**.



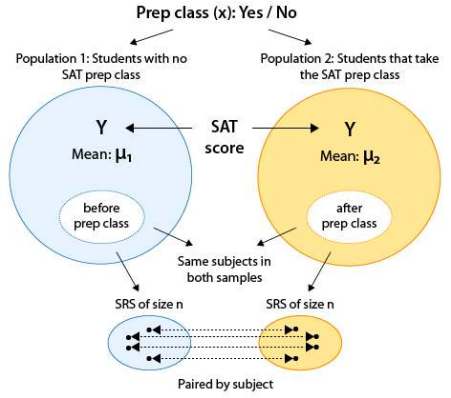
Notice from this point forward we will use the terms population 1 and population 2 instead of sub-population 1 and sub-population 2. Either terminology is correct.

One of the most common cases where dependent samples occur is when both samples have the same subjects and they are "**paired by subject**." In other words, **each subject is measured twice on the response variable**, typically **before** and then **after** some kind of treatment/intervention in order to assess its effectiveness.

EXAMPLE: SAT Prep Class

Suppose you want to assess the effectiveness of an SAT prep class.

It would make sense to use the matched pairs design and record each sampled student's SAT score before and after the SAT prep classes are attended:



Recall that the two populations represent the two values of the explanatory variable. In this situation, those two values come from **a single set of subjects**.

- In other words, both populations really have the **same students**.
- However, each population has a different value of the explanatory variable. Those values are: no prep class, prep class.

This, however, is not the only case where the paired design is used. Other cases are when the pairs are "**natural pairs**," such as **siblings**, **twins**, or **couples**.

Notes about graphical summaries for paired data in Case CQ:

- Due to the paired nature of this type of data, we cannot really use side-by-side boxplots to visualize this data as the information contained in the pairing is completely lost.
- We will need to provide graphical summaries of the differences themselves in order to explore this type of data.

The Idea Behind Paired t-Test

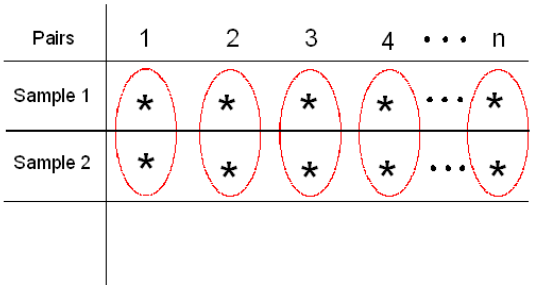
The idea behind the paired t-test is to **reduce** this **two-sample situation**, where we are comparing two means, **to a single sample situation** where we are doing inference on a single mean, and **then use a simple t-test** that we introduced in the previous module.

In this setting, we can easily reduce the raw data to a set of **differences** and conduct a **one-sample t-test.**

- Thus we simplify our inference procedure to a problem where we are making an inference about a single mean: **the mean of the differences**.

In other words, by **reducing the two samples to one sample of differences,** we are essentially **reducing the problem from a problem where we're comparing two means** (i.e., doing inference on $\mu_1 - \mu_2$) **to a problem in which we are studying one mean**.

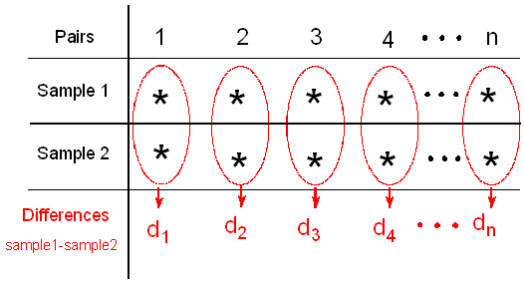
In general, in every matched pairs problem, **our data consist of 2 samples which are organized in n pairs**:



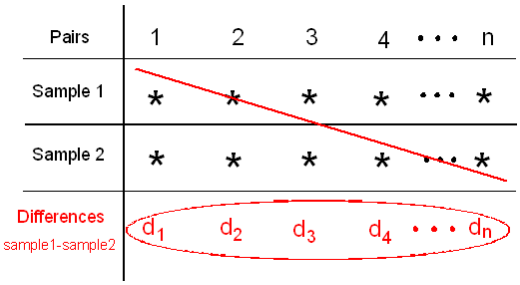
We reduce the two samples to only one by **calculating the difference between the two observations for each pair**.

For example, think of Sample 1 as "before" and Sample 2 as "after". We can find the difference between the before and after results for each participant, which gives us only one sample, namely "before – after". We label this difference as "d" in the illustration

below.



The **paired t-test is based on this one sample of n differences**,



and it **uses those differences as data for a one-sample t-test on a single mean** — the mean of the differences.

This is the general idea behind the paired t-test; it is nothing more than a regular one-sample t-test for the mean of the differences!

Test Procedure for Paired T-Test

We will now go through the 4-step process of the paired t-test.

- **Step 1: State the hypotheses**

Recall that in the t-test for a single mean our null hypothesis was: Ho: $\mu = \mu_0$ and the alternative was one of Ha: $\mu < \mu_0$ or $\mu > \mu_0$ or $\mu \neq \mu_0$. Since the paired t-test is a special case of the one-sample t-test, the hypotheses are the same except that:

> Instead of simply μ we use the notation μ_d to denote that the parameter of interest is the mean of the differences.
>
> In this course our null value μ_0 is always 0. In other words, going back to our original paired samples our null hypothesis claims that that there is no difference between the two means. (Technically, it does not have to be zero if you are interested in a more specific difference – for example, you might be interested in showing that there is a reduction in blood pressure of more than 10 points but we will not specifically look at such situations).

Therefore, in the paired t-test: The **null hypothesis** is always:

> **Ho: $\mu_d = 0$**
> (There IS NO association between the categorical explanatory variable and the quantitative response variable)

We will focus on the **two-sided alternative hypothesis** of the form:

> **Ha: $\mu_d \neq 0$**
> (There IS AN association between the categorical explanatory variable and the quantitative response variable)

Some students find it helpful to know that it turns out that $\mu_d = \mu_1 - \mu_2$ (in other words, the difference between the means is the same as the mean of the differences). You may find it easier to first think about the hypotheses in terms of $\mu_1 - \mu_2$ and then represent it in terms of μ_d.

Did I Get This? Setting up Hypotheses
(Non-Interactive Version – Spoiler Alert)

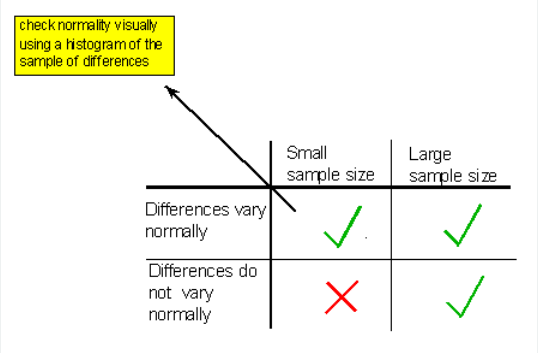
- **Step 2: Obtain data, check conditions, and summarize data**

The paired t-test, as a special case of a one-sample t-test, can be safely used as long as:

The sample of differences is **random** (or at least can be considered random in context).

The distribution of the differences in the population should vary normally if you have small samples. If the sample size is large, it is safe to use the paired t-test regardless of whether the differences vary normally or not. This condition is satisfied **in the three situations marked by a green check mark in the table below**.

Note: normality is checked by looking at the histogram of differences, and as long as no clear violation of normality (such as extreme skewness and/or outliers) is apparent, the normality assumption is reasonable.



Assuming that we can safely use the paired t-test, the data are summarized by a **test statistic:**

$$t = \frac{\bar{y}_d - 0}{s_d/\sqrt{n}}$$

where

$$\bar{y}_d = \text{ sample mean of the differences}$$

$$s_d = \text{sample standard deviation of the differences}$$

This **test statistic** measures (in standard errors) how far our data are (represented by the sample mean of the differences) from the null hypothesis (represented by the null value, 0).

Notice this test statistic has the same general form as those discussed earlier:

$$\text{test statistic} = \frac{\text{estimator - null value}}{\text{standard error of estimator}}$$

- **Step 3: Find the p-value of the test by using the test statistic as follows**

As a special case of the one-sample t-test, the **null distribution of the paired t-test statistic is a t distribution (with n – 1 degrees of freedom)**, which is the distribution under which the p-values are calculated. **We will use software to find the p-value for us.**

- **Step 4: Conclusion**

As usual, we draw our conclusion based on the p-value. Be sure to write your conclusions in context by specifying your current variables and/or precisely describing the population mean difference in terms of the current variables.

In particular, **if a cutoff probability, α (significance level), is specified, we reject Ho if the p-value is less than α. Otherwise, we fail to reject Ho.**

If the p-value is small, there is a statistically significant difference between what was observed in the sample and what was claimed in Ho, so we reject Ho.

Conclusion: There is enough evidence that the categorical explanatory variable is associated with the quantitative response variable. More specifically, there is enough evidence that the population mean difference is not equal to zero.

Remember: a small p-value tells us that there is very little chance of getting data like those observed (or even more extreme) if the null hypothesis were true. Therefore, a small p-value indicates that we should reject the null hypothesis.

If the p-value is not small, we do not have enough statistical evidence to reject Ho.

Conclusion: There is NOT enough evidence that the categorical explanatory variable is associated with the quantitative response variable. More specifically, there is NOT enough evidence that the population mean difference is not equal to zero.

Notice how much better the first sentence sounds! It can get difficult to correctly phrase these conclusions in terms of the mean difference without confusing double negatives.

Learning Objectives

LO 4.40: Based upon the output for a paired t-test, correctly interpret in context the appropriate confidence interval for the population mean-difference.

As in previous methods, we can **follow-up with a confidence interval for the mean difference, μ_d** and **interpret this interval in the context** of the problem.

Interpretation: We are 95% confident that the population mean difference (described in context) is between (lower bound) and (upper bound).

Confidence intervals can also be used to determine whether or not to reject the null hypothesis of the test based upon whether or not the null value of zero falls outside the interval or inside.

If the null value, 0, falls **outside** the confidence interval, **Ho is rejected**. (Zero is NOT a plausible value based upon the confidence interval)

If the null value, 0, falls **inside** the confidence interval, **Ho is not rejected**. (Zero IS a plausible value based upon the confidence interval)

NOTE: Be careful to choose the correct confidence interval about the population mean difference and not the individual confidence intervals for the means in the groups themselves.

Now let's look at an example.

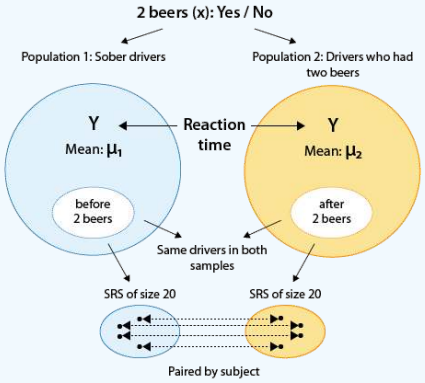
EXAMPLE: Drinking and Driving

Note: In some of the videos presented in the course materials, we do conduct the one-sided test for this data instead of the two-sided test we conduct below. In Unit 4B we are going to restrict our attention to two-sided tests supplemented by confidence intervals as needed to provide more information about the effect of interest.

- **Here is the SPSS Output for this example as well as the SAS Output and SAS Code.**

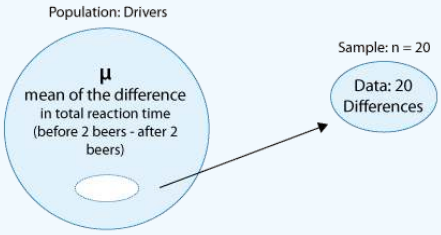
Drunk driving is one of the main causes of car accidents. Interviews with drunk drivers who were involved in accidents and survived revealed that one of the main problems is that drivers do not realize that they are impaired, thinking "I only had 1-2 drinks … I am OK to drive."

A sample of 20 drivers was chosen, and their reaction times in an obstacle course were measured before and after drinking two beers. The purpose of this study was to check whether drivers are impaired after drinking two beers. Here is a figure summarizing this study:



- Note that the **categorical explanatory variable here is "drinking 2 beers (Yes/No)"**, and the **quantitative response variable is the reaction time**.
- By using the matched pairs design in this study (i.e., by measuring each driver twice), the researchers isolated the effect of the two beers on the drivers and eliminated any other confounding factors that might influence the reaction times (such as the driver's experience, age, etc.).
- **For each driver, the two measurements are the total reaction time before drinking two beers, and after.** You can see the data by following the links in Step 2 below.

Since the measurements are paired, we can easily reduce the raw data to a set of **differences** and conduct a one-sample t-test.



Here are some of the results for this data:

Driver	1	2	3	4	...	20
Sample 1 (Before)	6.25	2.96	4.95	3.94	...	4.69
Sample 2 (After)	6.85	4.78	5.57	4.01	...	3.72
Differences (Before - After)	-0.60	-1.82	-0.62	-0.07		0.97

Step 1: State the hypotheses

We define μ_d = the population mean difference in reaction times (Before – After).

As we mentioned, the null hypothesis is:

- **Ho: $\mu_d = 0$** (indicating that the population of the differences are centered at a number that IS ZERO)

The null hypothesis claims that the differences in reaction times are centered at (or around) 0, indicating that drinking two beers has no real impact on reaction times. In other words, drivers are not impaired after drinking two beers.

Although we really want to know whether their reaction times are longer after the two beers, **we will still focus on conducting two-sided hypothesis tests**. We will be able to address whether the reaction times are longer after two beers when we look at the **confidence interval**.

Therefore, we will use the two-sided alternative:

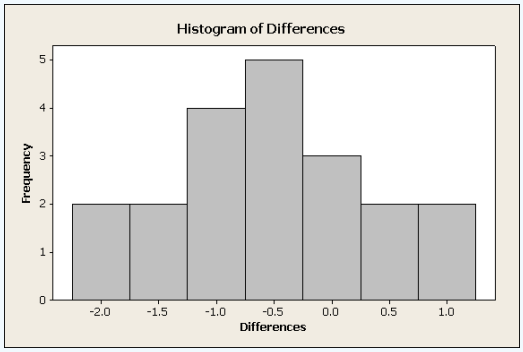
- **Ha: $\mu_d \neq 0$** (indicating that the population of the differences are centered at a number that is NOT ZERO)

Step 2: Obtain data, check conditions, and summarize data

- **Data:** Beers SPSS format, SAS format, Excel format, CSV format

Let's first check whether we can safely proceed with the paired t-test, by checking the two conditions.

- The sample of drivers was chosen at **random**.
- The **sample size is not large** (n = 20), so in order to proceed, we need to look at the histogram or QQ-plot of the differences and make sure there is no evidence that the normality assumption is not met.



We can see from the histogram above that there is no evidence of violation of the normality assumption (on the contrary, the histogram looks quite normal).

Also note that the vast majority of the differences are negative (i.e., the total reaction times for most of the drivers are larger after the two beers), suggesting that the data provide evidence against the null hypothesis.

The question (which the p-value will answer) is whether these data provide strong enough evidence or not against the null hypothesis. We can safely proceed to calculate the test statistic (which in practice we leave to the software to calculate for us).

Test Statistic: We will use software to calculate the **test statistic** which is **t = -2.58**.

- Recall: This indicates that the data (represented by the sample mean of the differences) are **2.58 standard errors below the null hypothesis** (represented by the null value, 0).

Step 3: Find the p-value of the test by using the test statistic as follows

As a special case of the one-sample t-test, the **null distribution of the paired t-test statistic is a t distribution (with n – 1 degrees of freedom)**, which is the distribution under which the p-values are calculated.

We will let the software find the p-value for us, and in this case, gives us a **p-value of 0.0183 (SAS) or 0.018 (SPSS).**

The small p-value tells us that there is very little chance of getting data like those observed (or even more extreme) if the null hypothesis were true. More specifically, there is less than a 2% chance (0.018=1.8%) of obtaining a test statistic of -2.58 (or lower) or 2.58 (or higher), assuming that 2 beers have no impact on reaction times.

Step 4: Conclusion

In our example, the p-value is 0.018, indicating that the data provide enough evidence to reject Ho.

- **Conclusion: There is enough evidence that drinking two beers is associated with differences in reaction times of drivers.**

Follow-up Confidence Interval:

As a follow-up to this conclusion, we quantify the effect that two beers have on the driver, using the 95% confidence interval for μ_d.

Using statistical software, we find that the 95% confidence interval for μ_d, the mean of the differences (before – after), is roughly **(-0.9, -0.1).**

Note: Since the differences were calculated before-after, longer reaction times after the beers would translate into negative differences.

- **Interpretation: We are 95% confident that after drinking two beers, the true mean increase in total reaction time of drivers is between 0.1 and 0.9 of a second.**
- Thus, the results of the study do indicate impairment of drivers (longer reaction times) not the other way around!

Since the confidence interval does not contain the null value of zero, we can use it to decide to reject the null hypothesis. Zero is not a plausible value of the population mean difference based upon the confidence interval. Notice that using this method is not always practical as often we still need to provide the p-value in clinical research. (**Note:** this is NOT the interpretation of the confidence interval but a method of using the confidence interval to conduct a hypothesis test.)

Did I Get This? Confidence Intervals for the Population Mean Difference
(Non-Interactive Version – Spoiler Alert)

Practical Significance:

We should definitely ask ourselves if this is practically significant and I would argue that it is.

- Although a difference in the mean reaction time of 0.1 second might not be too bad, a difference of 0.9 seconds is likely a problem.
- Even at a difference in reaction time of 0.4 seconds, if you were traveling 60 miles per hour, this would translate into a distance traveled of around 35 feet.

Many Students Wonder: One-sided vs. Two-sided P-values

In the output, we are generally provided the two-sided p-value. We must be very careful when converting this to a one-sided p-value (if this is not provided by the software)

- **IF the data are in the direction of our alternative hypothesis** then we can simply take **half of the two-sided p-value.**
- **IF, however, the data are NOT in the direction of the alternative**, the correct p-value is VERY LARGE and is the **complement of (one minus) half the two-sided p-value.**

The "driving after having 2 beers" example is a case in which observations are paired by subject. In other words, both samples have the same subject, so that each subject is measured twice. Typically, as in our example, one of the measurements occurs before a treatment/intervention (2 beers in our case), and the other measurement after the treatment/intervention.

Our next example is another typical type of study where the matched pairs design is used—it is a study involving twins.

✔ EXAMPLE: IQ Scores

Researchers have long been interested in the extent to which **intelligence, as measured by IQ score, is affected by "nurture" as opposed to "nature"**: that is, are people's IQ scores mainly a result of their upbringing and environment, or are they mainly an inherited trait?

A study was designed to measure the effect of home environment on intelligence, or more specifically, the study was designed to address the question: "Are there statistically significant differences in IQ scores between people who were raised by their birth parents, and those who were raised by someone else?"

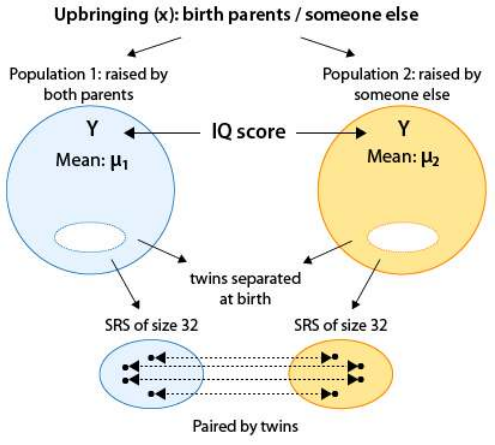
In order to be able to answer this question, the researchers needed to get two groups of subjects (one from the population of people who were raised by their birth parents, and one from the population of people who were raised by someone else) who are as similar as possible in all other respects. In particular, since genetic differences may also affect intelligence, the researchers wanted to control for this confounding factor.

We know from our discussion on study design (in the Producing Data unit of the course) that one way to (at least theoretically) control for all confounding factors is randomization—randomizing subjects to the different treatment groups. In this case, however, this is not possible. This is an observational study; you cannot randomize children to either be raised by their birth parents or to be raised by someone else. How else can we eliminate the genetics factor? We can conduct a "twin study."

Because identical twins are genetically the same, a good design for obtaining information to answer this question would be to compare IQ scores for identical twins, one of whom is raised by birth parents and the other by someone else. Such a design (matched pairs) is an excellent way of making a comparison between individuals who only differ with respect to the explanatory variable of interest (upbringing) but are as alike as they can possibly be in all other important aspects (inborn intelligence). Identical twins raised apart were studied by Susan Farber, who published her studies in the book "Identical Twins Reared Apart" (1981, Basic Books).

In this problem, we are going to use the data that appear in Farber's book in table E6, of the IQ scores of 32 pairs of identical twins who were reared apart.

Here is a figure that will help you understand this study:



Here are the important things to note in the figure:

- We are essentially **comparing** the **mean IQ scores in two populations** that are **defined by** our (two-valued categorical) **explanatory variable** — **upbringing** (X), whose two values are: **raised by birth parents, raised by someone else**.
- This is a **matched pairs design** (as opposed to a two independent samples design), since each observation in one sample is **linked (matched)** with an observation in the second sample. The observations are paired by twins.

Each of the 32 rows represents one pair of twins. Keeping the notation that we used above, twin 1 is the twin that was raised by his/her birth parents, and twin 2 is the twin that was raised by someone else. Let's carry out the analysis.

Step 1: State the hypotheses

Recall that in matched pairs, we reduce the data from two samples to one sample of differences:

Pair	1	2	3	4	•••	32
TWIN 1 (birth parents)	113	94	99	77	•••	97
TWIN 2 (someone else)	109	100	86	80	•••	98
Differences (twin 1 - twin 2)	4	-6	13	-3	•••	-1

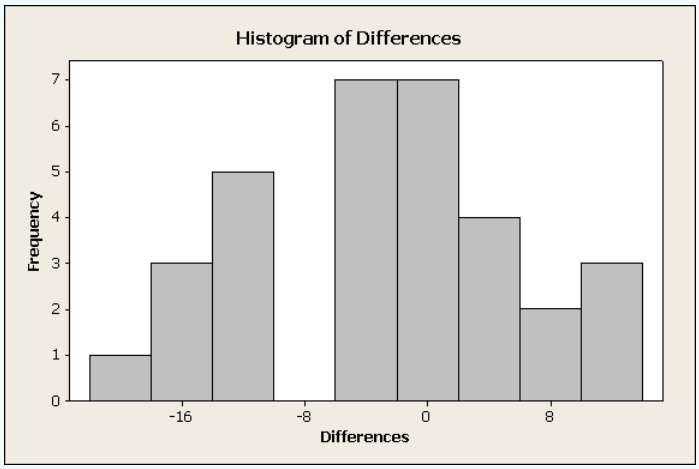
The hypotheses are stated in terms of the mean of the difference where, μ_d = population mean difference in IQ scores (Birth Parents – Someone Else):

- **Ho:** $\mu_d = 0$ (indicating that the population of the differences are centered at a number that IS ZERO)
- **Ha:** $\mu_d \neq 0$ (indicating that the population of the differences are centered at a number that is NOT ZERO)

Step 2: Obtain data, check conditions, and summarize data

Is it safe to use the paired t-test in this case?

- Clearly, the samples of twins are not random samples from the two populations. However, in this context, they can be considered as random, assuming that there is nothing special about the IQ of a person just because he/she has an identical twin.
- The sample size here is n = 32. Even though it's the case that if we use the n > 30 rule of thumb our sample can be considered large, it is sort of a borderline case, so just to be on the safe side, we should look at the histogram of the differences just to make sure that we do not see anything extreme. (Comment: Looking at the histogram of differences in every case is useful even if the sample is very large, just in order to get a sense of the data. Recall: "Always look at the data.")



The data don't reveal anything that we should be worried about (like very extreme skewness or outliers), so we can safely proceed. Looking at the histogram, we note that most of the differences are negative, indicating that in most of the 32 pairs of twins, twin 2 (raised by someone else) has a higher IQ.

From this point we rely on statistical software, and find that:

- **t-value = -1.85**
- **p-value = 0.074**

Our test statistic is -1.85.

Our data (represented by the sample mean of the differences) are 1.85 standard errors below the null hypothesis (represented by the null value 0).

Step 3: Find the p-value of the test by using the test statistic as follows

The p-value is 0.074, indicating that there is a 7.4% chance of obtaining data like those observed (or even more extreme) assuming that H_o is true (i.e., assuming that there are no differences in IQ scores between people who were raised by their natural parents and those who weren't).

Step 4: Conclusion

Using the conventional significance level (cut-off probability) of .05, our p-value is not small enough, and we therefore cannot reject H_o.

- **Conclusion:** Our data do not provide enough evidence to conclude that whether a person was raised by his/her natural parents has an impact on the person's intelligence (as measured by IQ scores).

Confidence Interval:

The 95% confidence interval for the population mean difference is (-6.11322, 0.30072).

Interpretation:

- **We are 95% confident that the population mean IQ for twins raised by someone else is between 6.11 greater to 0.3 lower than that for twins raised by their birth parents.**

- **OR … We are 95% confident that the population mean IQ for twins raised by their birth parents is between 6.11 lower to 0.3 greater than that for twins raised by someone else.**
- **Note:** The order of the groups as well as the numbers provided in the interval can vary, what is important is to get the "lower" and "greater" with the correct value based upon the group order being used.
  - Here we used Birth Parents – Someone Else and thus a positive number for our population mean difference indicates that birth parents group is higher (someone else gorup is lower) and a negative number indicates the someone else group is higher (birth parents group is lower).

This confidence interval does contain zero and thus results in the same conclusion to the hypothesis test. Zero IS a plausible value of the population mean difference and thus we cannot reject the null hypothesis.

Practical Significance:

- The confidence interval does "lean" towards the difference being negative, indicating that in most of the 32 pairs of twins, twin 2 (raised by someone else) has a higher IQ. The sample mean difference is -2.9 so we would need to consider whether this value and range of plausible values have any real practical significance.
- In this case, I don't think I would consider a difference in IQ score of around 3 points to be very important in practice (but others could reasonably disagree).

It is very important to pay attention to whether the two-sample t-test or the paired t-test is appropriate. In other words, being aware of the study design is extremely important. Consider our example, if we had not "caught" that this is a matched pairs design, and had analyzed the data as if the two samples were independent using the two-sample t-test, we would have obtained a p-value of 0.114.

Note that using this (wrong) method to analyze the data, and a significance level of 0.05, we would conclude that the data do not provide enough evidence for us to conclude that reaction times differed after drinking two beers. This is an example of how using the wrong statistical method can lead you to wrong conclusions, which in this context can have very serious implications.

Comments:

- The 95% confidence interval for μ can be used here in the same way as for proportions to conduct the two-sided test (checking whether the null value falls inside or outside the confidence interval) or following a t-test where Ho was rejected to get insight into the value of μ.
- In most situations in practice we use two-sided hypothesis tests, followed by confidence intervals to gain more insight.

Now try a complete example for yourself.

Learn By Doing: Matched Pairs – Gosset's Seed Data
(Non-Interactive Version – Spoiler Alert)

Additional Data for Practice

Here are two other datasets with paired samples.

- Seeds: SPSS format, SAS format, Excel format, CSV format
- Twins: SPSS format, SAS format, Excel format, CSV format

Non-Parametric Alternatives for Matched Pair Data

Learning Objectives

LO 5.1: For a data analysis situation involving two variables, determine the appropriate alternative (non-parametric) method when assumptions of our standard methods are not met.

The statistical tests we have previously discussed (and many we will discuss) require assumptions about the distribution in the population or about the requirements to use a certain approximation as the sampling distribution. These methods are called **parametric**.

When these assumptions are not valid, alternative methods often exist to test similar hypotheses. Tests which require only minimal distributional assumptions, if any, are called **non-parametric** or **distribution-free** tests.

At the end of this section we will provide some details (see Details for Non-Parametric Alternatives), for now we simply want to mention that there are **two common non-parametric alternatives to the paired t-test**. They are:

- **Sign** Test
- Wilcoxon **Sign**ed-Rank Test

The fact that both of these tests have the word "**sign**" in them is not a coincidence – it is due to the fact that we will be interested in whether the differences have a positive **sign** or a negative **sign** – and the fact that this word appears in both of these tests can help you to remember that they correspond to **paired methods** where we are often interested in whether there was an increase (positive **sign**) or a decrease (negative **sign**).

Let's Summarize

- The **paired t-test** is used to compare **two population means** when the two samples (drawn from the two populations) are **dependent** in the sense that every observation in one sample can be **linked** to an observation in the other sample. Such a design is called "**matched pairs**."
- The most common case in which the matched pairs design is used is when the **same subjects** are **measured twice**, usually before and then after some kind of treatment and/or intervention. Another classic case are studies involving twins.
- In the background, we have a **two-valued categorical explanatory** whose **categories define** the **two populations we are comparing** and whose effect on the response variable we are trying to assess.
- The **idea** behind the paired t-test is to **reduce the data from two samples to just one sample of the differences**, and use these observed differences as data for **inference about a single mean** — the mean of the differences, μ_d.
- The paired t-test is therefore simply a **one-sample t-test for the mean of the differences μ_d**, where the **null value is 0**.
- Once we verify that we can safely proceed with the paired t-test, **we use software output to carry it out**.
- A **95% confidence interval for μ_d** can be very **insightful** after a test has rejected the null hypothesis, and can also be used for testing in the two-sided case.
- Two **non-parametric alternatives** to the paired t-test are the **sign test** and the **Wilcoxon signed–rank test**. (See Details for Non-Parametric Alternatives.)

Two Independent Samples

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results

Learning Objectives

LO 4.35: For a data analysis situation involving two variables, choose the appropriate inferential method for examining the relationship between the variables and justify the choice.

Learning Objectives

LO 4.36: For a data analysis situation involving two variables, carry out the appropriate inferential method for examining relationships between the variables and draw the correct conclusions in context.

CO-5: Determine preferred methodological alternatives to commonly used statistical methods when assumptions are not met.

REVIEW: Unit 1 Case C-Q

Video

Video: Two Independent Samples (38:56)

Related SAS Tutorials

- 7A (2:32) Numeric Summaries by Groups
- 7B (3:03) Side-By-Side Boxplots
- 7C (6:57) Two Sample T-Test

Related SPSS Tutorials

- 7A (3:29) Numeric Summaries by Groups
- 7B (1:59) Side-By-Side Boxplots
- 7C (5:30) Two Sample T-Test

Introduction

Here is a summary of the tests we will learn for the scenario where k = 2. Methods in **BOLD** will be our main focus.

We have completed our discussion on dependent samples (2nd column) and now we move on to independent samples (1st column).

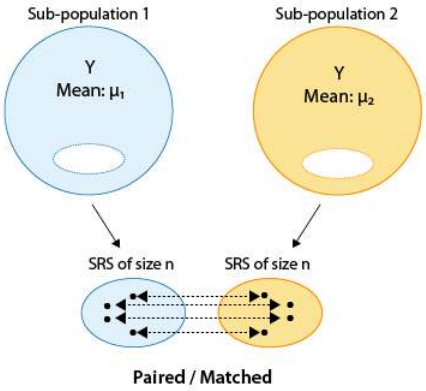
Independent Samples (More Emphasis)	Dependent Samples (Less Emphasis)
Standard Tests • **Two Sample T-Test Assuming Equal Variances** • **Two Sample T-Test Assuming Unequal Variances** Non-Parametric Test • Mann-Whitney U (or Wilcoxon Rank-Sum) Test	**Standard Test** • **Paired T-Test** Non-Parametric Tests • Sign Test • Wilcoxon Signed-Rank Test

Dependent vs. Independent Samples

Learning Objectives

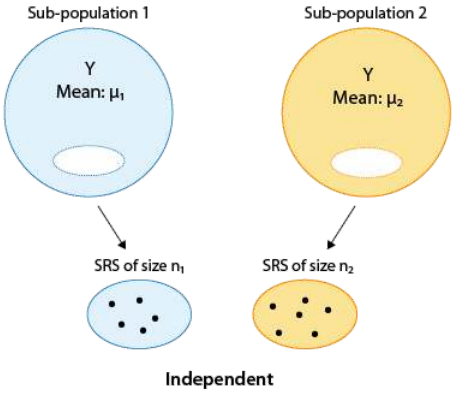
LO 4.37: Identify and distinguish between independent and dependent samples.

We have discussed the **dependent sample** case where observations are **matched/paired/linked** between the two samples. Recall that in that scenario observations can be the same individual or two individuals who are matched between samples. To analyze data from dependent samples, we simply took the differences and analyzed the difference using one-sample techniques.



Now we will discuss the independent sample case. In this case, all individuals are independent of all other individuals in their sample as well as all individuals in the other sample. This is most often accomplished by either:

- **Taking a random sample from each of the two groups under study.** For example to compare heights of males and females, we could take a random sample of 100 females and another random sample of 100 males. The result would be two samples which are independent of each other.
- **Taking a random sample from the entire population and then dividing it into two sub-samples based upon the grouping variable of interest.** For example, we take a random sample of U.S. adults and then split them into two samples based upon gender. This results in a sub-sample of females and a sub-sample of males which are independent of each other.



Comparing Two Means – Two Independent Samples T-test

Learning Objectives

LO 4.38: In a given context, determine the appropriate standard method for comparing groups and provide the correct conclusions given the appropriate software output.

Learning Objectives

LO 4.39: In a given context, set up the appropriate null and alternative hypotheses for comparing groups.

Recall that here we are interested in the effect of a **two-valued (k = 2) categorical variable** (X) on a **quantitative response** (Y). Random samples from the two sub-populations (defined by the two categories of X) are obtained and we need to evaluate whether or not the data provide enough evidence for us to believe that the two sub-population means are different.

In other words, our goal is to test whether the means μ_1 and μ_2 (which are the means of the variable of interest in the two sub-populations) are equal or not, and in order to do that we have two samples, one from each sub-population, which were chosen independently of each other.

The test that we will learn here is commonly known as the **two-sample t-test**. As the name suggests, this is a t-test, which as we know means that the p-values for this test are calculated under some t-distribution.

Here are figures that illustrate some of the examples we will cover. Notice how the original variables X (categorical variable with two levels) and Y (quantitative variable) are represented. Think about the fact that we are in case C → Q!

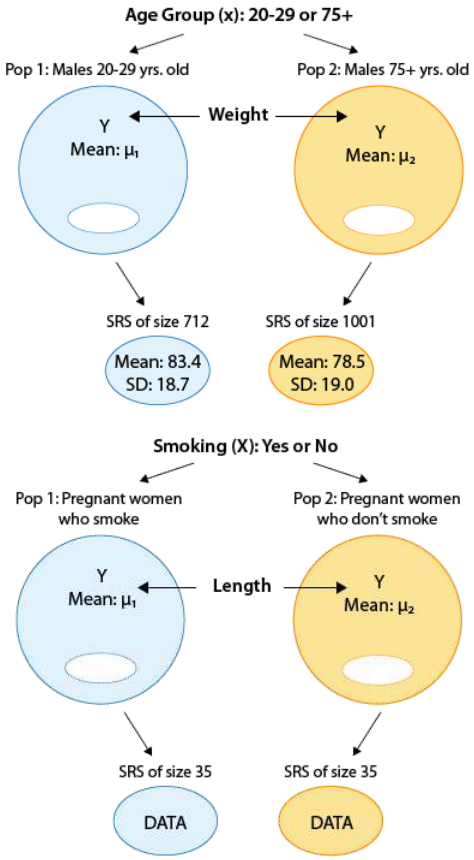
As in our discussion of dependent samples, we will often simplify our terminology and simply use the terms “population 1” and “population 2” instead of referring to these as sub-populations. Either terminology is fine.

Many Students Wonder: Two Independent Samples

Question: Does it matter which population we label as population 1 and which as population 2?

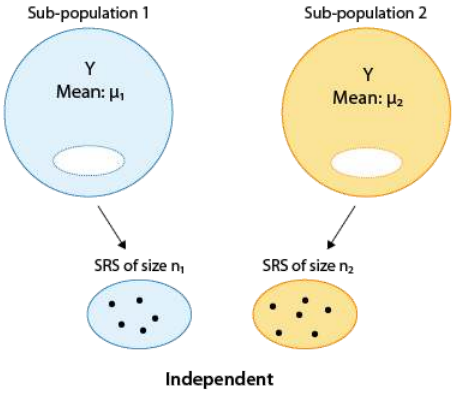
Answer: No, it does not matter as long as you are **consistent**, meaning that you do not switch labels in the middle.

- **BUT… considering how you label the populations is important in stating the hypotheses and in the interpretation of the results.**



Steps for the Two-Sample T-test

Recall that our goal is to compare the means μ_1 and μ_2 based on the two independent samples.



- **Step 1: State the hypotheses**

The hypotheses represent our goal to compare $\mu_1$and μ_2.

The **null hypothesis is always:**

Ho: $\mu_1 - \mu_2 = 0$ (which is the same as $\mu_1 = \mu_2$)
(There IS NO association between the categorical explanatory variable and the quantitative response variable)

We will focus on the **two-sided alternative hypothesis** of the form:

Ha: $\mu_1 - \mu_2 \neq 0$ (which is the same as $\mu_1 \neq \mu_2$) **(two-sided)**
(There IS AN association between the categorical explanatory variable and the quantitative response variable)

Note that the null hypothesis claims that there is no difference between the means. Conceptually, Ho claims that there is no relationship between the two relevant variables (X and Y).

Our parameter of interest in this case (the parameter about which we are making an inference) is the difference between the means ($\mu_1 - \mu_2$) and the null value is 0. The alternative hypothesis claims that there is a difference between the means.

Did I Get This? What do our hypotheses mean in context?
(Non-Interactive Version – Spoiler Alert)

- **Step 2: Obtain data, check conditions, and summarize data**

The two-sample t-test can be safely used as long as the following conditions are met:

The two samples are indeed independent.

We are in one of the following two scenarios:

(i) Both populations are normal, or more specifically, the distribution of the response Y in both populations is normal, and both samples are random (or at least can be considered as such). In practice, checking normality in the populations is done by looking at each of the samples using a histogram and checking whether there are any signs that the populations are not normal. Such signs could be extreme skewness and/or extreme outliers.

(ii) The populations are known or discovered not to be normal, but the sample size of each of the random samples is large enough (we can use the rule of thumb that a sample size greater than 30 is considered large enough).

Did I Get This? Conditions for Two Independent Samples
(Non-Interactive Version – Spoiler Alert)

Assuming that we can safely use the two-sample t-test, we need to summarize the data, and in particular, calculate our data summary—the test statistic.

Test Statistic for Two-Sample T-test:

There are two choices for our test statistic, and **we must choose** the appropriate one to summarize our data We will see how to choose between the two test statistics in the next section. The two options are as follows:

We use the following notation to describe our samples:

n_1, n_2 = sample sizes of the samples from population 1 and population 2

$\bar{y}_1, \bar{y}_2$ = sample means of the samples from population 1 and population 2

s_1, s_2 = sample standard deviations of the samples from population 1 and population 2

s_p = pooled estimate of a common population standard deviation

Here are the two cases for our test statistic.

(A) Equal Variances: If it is safe to assume that the **two populations have equal standard deviations**, we can pool our estimates of this common population standard deviation and use the following test statistic.

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

(B) Unequal Variances: If it is NOT safe to assume that the two populations have equal standard deviations, we have **unequal standard deviations** and must use the following test statistic.

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Comments:

- It is possible to never assume equal variances; however, if the assumption of equal variances is satisfied the equal variances t-test will have greater power to detect the difference of interest.
- We will not be calculating the values of these test statistics by hand in this course. We will instead rely on software to obtain the value for us.
- Both of these test statistics measure (in standard errors) how far our data are (represented by the difference of the sample means) from the null hypothesis (represented by the null value, 0).
- These test statistics have the same general form as others we have discussed. We will not discuss the derivation of the standard errors in each case but you should understand this general form and be able to identify each component for a specific test statistic.

$$\text{test statistic} = \frac{\text{estimator - null value}}{\text{standard error of estimator}}$$

- **Step 3: Find the p-value of the test by using the test statistic as follows**

Each of these tests rely on a particular t-distribution under which the p-values are calculated. In the case where equal variances are assumed, the degrees of freedom are simply:

$$n_1 + n_2 - 2$$

whereas in the case of unequal variances, the formula for the degrees of freedom is more complex. We will rely on the software to obtain the degrees of freedom in both cases and provided us with the correct p-value (usually this will be a two-sided p-value).

- **Step 4: Conclusion**

As usual, we draw our conclusion based on the p-value. Be sure to write your conclusions in context by specifying your current variables and/or precisely describing the difference in population means in terms of the current variables.

> **If the p-value is small**, there is a statistically significant difference between what was observed in the sample and what was claimed in Ho, so we reject Ho.
>
> > **Conclusion:** There is enough evidence that the categorical explanatory variable is related to (or associated with) the quantitative response variable. More specifically, there is enough evidence that the difference in population means is not equal to zero.
>
> **If the p-value is not small**, we do not have enough statistical evidence to reject Ho.
>
> > **Conclusion:** There is NOT enough evidence that the categorical explanatory variable is related to (or associated with) the quantitative response variable. More specifically, there is enough evidence that the difference in population means is not equal to zero.
>
> In particular, if a cutoff probability, α (significance level), is specified, we reject Ho if the p-value is less than α. Otherwise, we do not reject Ho.

Learning Objectives

LO 4.41: Based upon the output for a two-sample t-test, correctly interpret in context the appropriate confidence interval for the difference between population means

As in previous methods, we can **follow-up with a confidence interval for the difference between population means, $\mu_1 - \mu_2$** and **interpret this interval in the context** of the problem.

> **Interpretation:** We are 95% confident that the population mean for (one group) is between _____ compared to the population mean for (the other group).

Confidence intervals can also be used to determine whether or not to reject the null hypothesis of the test based upon whether or not the null value of zero falls outside the interval or inside.

> If the null value, 0, falls **outside** the confidence interval, **Ho is rejected**. (Zero is NOT a plausible value based upon the confidence interval)

If the null value, 0, falls **inside** the confidence interval, **Ho is not rejected**. (Zero IS a plausible value based upon the confidence interval)

NOTE: Be careful to choose the correct confidence interval about the difference between population means using the same assumption (variances equal or variances unequal) and not the individual confidence intervals for the means in the groups themselves.

Many Students Wonder: Reading Statistical Software Output for Two-Sample T-test

Test for Equality of Variances (or Standard Deviations)

Learning Objectives

LO 4.42: Based upon the output for a two-sample t-test, determine whether to use the results assuming equal variances or those assuming unequal variances.

Since we have two possible tests we can conduct, based upon whether or not we can assume the population standard deviations (or variances) are equal, we need a method to determine which test to use.

Although you can make a reasonable guess using information from the data (i.e. look at the distributions and estimates of the standard deviations and see if you feel they are reasonably equal), we have a test which can help us here, called the **test for Equality of Variances**. This output is automatically displayed in many software packages when a two-sample t-test is requested although the particular test used may vary.The hypotheses of this test are:

Ho: $\sigma_1 = \sigma_2$ (the standard deviations in the two populations are the same)

Ha: $\sigma_1 \neq \sigma_2$ (the standard deviations in the two populations are not the same)

- **If the p-value of this test for equal variances is small**, there is enough evidence that the standard deviations in the two populations are different and **we cannot assume equal variances**.
  - **IMPORTANT! In this case, when we conduct the two-sample t-test to compare the population means, we use the test statistic for unequal variances.**
- **If the p-value of this test is large**, there is not enough evidence that the standard deviations in the two populations are different. In this case **we will assume equal variances since we have no clear evidence to the contrary**.
  - **IMPORTANT! In this case, when we conduct the two-sample t-test to compare the population means, we use the test statistic for equal variances.**

Now let's look at a complete example of conducting a two-sample t-test, including the embedded test for equality of variances.

EXAMPLE: What is more important - personality or looks?

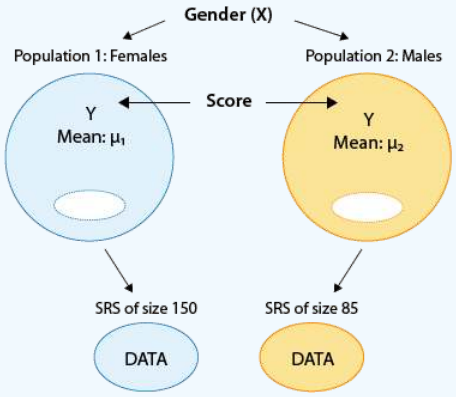
This question was asked of a random sample of 239 college students, who were to answer on a scale of 1 to 25. An answer of 1 means personality has maximum importance and looks no importance at all, whereas an answer of 25 means looks have maximum importance and personality no importance at all. The purpose of this survey was to examine whether males and females differ with respect to the importance of looks vs. personality.

Note that the data have the following format:

Score (Y)	Gender (X)
15	Male
13	Female
10	Female

Score (Y)	Gender (X)
12	Male
14	Female
14	Male
6	Male
17	Male
etc.	

The format of the data reminds us that we are essentially examining the relationship between the two-valued categorical variable, gender, and the quantitative response, score. The two values of the categorical explanatory variable (k = 2) define the two populations that we are comparing — males and females. The comparison is with respect to the response variable score. Here is a figure that summarizes the example:



Comments:

- Note that this figure emphasizes how the fact that our explanatory is a two-valued categorical variable means that in practice we are comparing two populations (defined by these two values) with respect to our response Y.
- Note that even though the problem description just says that we had 239 students, the figure tells us that there were 85 males in the sample, and 150 females.
- Following up on comment 2, note that 85 + 150 = 235 and not 239. In these data (which are real) there are four "missing observations," 4 students for which we do not have the value of the response variable, "importance." This could be due to a number of reasons, such as recording error or non response. The bottom line is that even though data were collected from 239 students, effectively we have data from only 235. (Recommended: Go through the data file and note that there are 4 cases of missing observations: students 34, 138, 179, and 183).

Step 1: State the hypotheses

Recall that the purpose of this survey was to examine whether the opinions of females and males **differ** with respect to the importance of looks vs. personality. The hypotheses in this case are therefore:

Ho: $\mu_1 - \mu_2 = 0$ (which is the same as $\mu_1 = \mu_2$)

Ha: $\mu_1 - \mu_2 \neq 0$ (which is the same as $\mu_1 \neq \mu_2$)

where μ_1 represents the mean "looks vs personality score" for females and μ_2 represents the mean "looks vs personality score" for males.

It is important to understand that conceptually, the two hypotheses claim:

Ho: Score (of looks vs. personality) is not related to gender

Ha: Score (of looks vs. personality) is related to gender

Step 2: Obtain data, check conditions, and summarize data

- **Data:** Looks SPSS format, SAS format, Excel format, CSV format
- Let's first check whether the conditions that allow us to safely use the two-sample t-test are met.
  - Here, 239 students were chosen and were naturally divided into a sample of females and a sample of males. Since the students were chosen at random, **the sample of females is independent of the sample of males.**
  - Here we are in the second scenario — **the sample sizes (150 and 85), are definitely large enough**, and so we can proceed regardless of whether the populations are normal or not.
- In the output below we first look at the **test for equality of variances (outlined in orange)**. The **two-sample t-test results we will use are outlined in blue**.
- There are TWO TESTS represented in this output and we must make the correct decision for BOTH of these tests to correctly proceed.
- **SOFTWARE OUTPUT In SPSS:**
  - The p-value for the test of equality of variances is reported as **0.849** in the **SIG column under Levene's test for equality of variances**. (Note this differs from the p-value found using SAS, two different tests are used by default between the two programs).
  - So we fail to reject the null hypothesis that the variances, or equivalently the standard deviations, are equal (**Ho: $\sigma_1 = \sigma_2$**).
  - **Conclusion to test for equality of variances:** We cannot conclude there is a difference in the variance of looks vs. personality score between males and females.
  - This results in using the row for Equal variances assumed to find the t-test results including the test statistic, p-value, and confidence interval for the difference. (Outlined in **BLUE**)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence ... Lower	95% Confidence ... Upper
Score (Y)	Equal variances assumed	.036	.849	-4.584	233	.000	-2.596	.566	-3.712	-1.480
	Equal variances not assumed			-4.657	182.973	.000	-2.596	.557	-3.696	-1.496

The output might also be broken up if you export or copy the items in certain ways. The results are the same but it can be more difficult to read.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
Score (Y)	Equal variances assumed	.036	.849	-4.584	233
	Equal variances not assumed			-4.657	182.973

Independent Samples Test

		t-test for Equality of Means			
		Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence ... Lower
Score (Y)	Equal variances assumed	.000	-2.596	.566	-3.712
	Equal variances not assumed	.000	-2.596	.557	-3.696

Independent Samples Test

		t-test for Equality of ... 95% Confidence ... Upper
Score (Y)	Equal variances assumed	-1.480
	Equal variances not assumed	-1.496

- **SOFTWARE OUTPUT In SAS:**
  - The p-value for the test of equality of variances is reported as **0.5698 in the Pr > F column under equality of variances**. (Note this differs from the p-value found using SPSS, two different tests are used by default between the two programs).
  - So we fail to reject the null hypothesis that the variances, or equivalently the standard deviations, are equal (**Ho: $\sigma_1 = \sigma_2$**).
  - **Conclusion to test for equality of variances:** We cannot conclude there is a difference in the variance of looks vs. personality score between males and females.
  - This results in using the row for POOLED method where equal variances are assumed to find the t-test results including the test statistic, p-value, and confidence interval for the difference. (Outlined in **BLUE**)

GenderX	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		10.7333	10.0469	11.4198	4.2548	3.8216	4.7995
Male		13.3294	12.4625	14.1963	4.0190	3.4924	4.7341
Diff (1-2)	Pooled	-2.5961	-3.7118	-1.4804	4.1713	3.8245	4.5878
Diff (1-2)	Satterthwaite	-2.5961	-3.6959	-1.4963			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	233	-4.58	<.0001
Satterthwaite	Unequal	182.97	-4.66	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	149	84	1.12	0.5698

- **TEST STATISTIC for Two-Sample T-test:** In all of the results above, we determine that we will use the test which assumes the variances are EQUAL, and we find our **test statistic** of **t = -4.58.**

Step 3: Find the p-value of the test by using the test statistic as follows

- We will let the software find the p-value for us, and in this case, **the p-value is less than our significance level of 0.05 in fact it is practically 0.**
- This is found in **SPSS in the equal variances assumed row under t-test in the SIG. (two-tailed) column given as 0.000** and in **SAS in the POOLED ROW under Pr > |t| column given as <0.0001.**
- A p-value which is practically 0 means that it would be almost impossible to get data like that observed (or even more extreme) had the null hypothesis been true.
- More specifically, in our example, if there were no differences between females and males with respect to whether they value looks vs. personality, it would be almost impossible (probability approximately 0) to get data where the difference between the sample means of females and males is -2.6 (that difference is 10.73 – 13.33 = -2.6) or more extreme.
- **Comment:** Note that the output tells us that the difference $\mu_1 - \mu_2$ is approximately -2.6. But more importantly, we want to know if this difference is statistically significant. To answer this, we use the fact that this difference is 4.58 standard errors below the null value.

Step 4: Conclusion

As usual a small p-value provides evidence against Ho. In our case our p-value is practically 0 (which is smaller than any level of significance that we will choose). The data therefore provide very strong evidence against Ho so we reject it.

- **Conclusion: There is enough evidence that the mean Importance score (of looks vs personality) of males differs from that of females. In other words, males and females differ with respect to how they value looks vs. personality.**

As a follow-up to this conclusion, we can construct a confidence interval for the difference between population means. In this case we will construct a confidence interval for $\mu_1 - \mu_2$ the population mean "looks vs personality score" for females minus the population mean "looks vs personality score" for males.

- Using statistical software, we find that the 95% confidence interval for $\mu_1 - \mu_2$ is roughly **(-3.7, -1.5).**
- This is found in **SPSS in the equal variances assumed row under 95% confidence interval columns given as -3.712 to -1.480** and in **SAS in the POOLED ROW under 95% CL MEAN column given as -3.7118 to -1.4804** (be careful NOT to choose the confidence interval for the standard deviation in the last column, 9% CL Std Dev).
- **Interpretation:**
  - **We are 95% confident that the population mean "looks vs personality score" for females is between 3.7 and 1.5 points lower than that of males.**
  - OR
  - **We are 95% confident that the population mean "looks vs personality score" for males is between 3.7 and 1.5 points higher than that of females.**
- The confidence interval therefore quantifies the effect that the explanatory variable (gender) has on the response (looks vs personality score).
- Since low values correspond to personality being more important and high values correspond to looks being more important, the result of our investigation suggests that, on average, females place personality higher than do males. Alternatively we could say that males place looks higher than do females.
- **Note:** The confidence interval does not contain zero (both values are negative based upon how we chose our groups) and thus using the confidence interval we can reject the null hypothesis here.

Practical Significance:

We should definitely ask ourselves if this is practically significant

- Is a true difference in population means as represented by our estimate from this data meaningful here? I will let you consider and answer for yourself.

SPSS Output for this example (Non-Parametric Output for Examples 1 and 2)

SAS Output and SAS Code (Includes Non-Parametric Test)

Here is another example.

EXAMPLE: BMI vs. Gender in Heart Attack Patients

A study was conducted which enrolled and followed heart attack patients in a certain metropolitan area. In this example we are interested in determining if there is a relationship between Body Mass Index (BMI) and gender. Individuals presenting to the hospital with a heart attack were randomly selected to participate in the study.

Step 1: State the hypotheses

Ho: $\mu_1 - \mu_2 = 0$ (which is the same as $\mu_1 = \mu_2$)

Ha: $\mu_1 - \mu_2 \neq 0$ (which is the same as $\mu_1 \neq \mu_2$)

where μ_1 represents the mean BMI for males and μ_2 represents the mean BMI for females.

It is important to understand that conceptually, the two hypotheses claim:

Ho: BMI is not related to gender in heart attack patients

Ha: BMI is related to gender in heart attack patients

Step 2: Obtain data, check conditions, and summarize data

- Data: WHAS500 SPSS format, SAS format
- Let's first check whether the conditions that allow us to safely use the two-sample t-test are met.
  - Here, subjects were chosen and were naturally divided into a sample of females and a sample of males. Since the subjects were chosen at random, the sample of females is independent of the sample of males.
  - Here, we are in the second scenario — the sample sizes are extremely large, and so we can proceed regardless of whether the populations are normal or not.
- In the output below we first look at the **test for equality of variances (outlined in orange)**. The **two-sample t-test results we will use are outlined in blue**.
- There are TWO TESTS represented in this output and we must make the correct decision for BOTH of these tests to correctly proceed.
- **SOFTWARE OUTPUT In SPSS:**
  - The p-value for the test of equality of variances is reported as **0.001** in the **SIG column under Levene's test for equality of variances**.
  - So we reject the null hypothesis that the variances, or equivalently the standard deviations, are equal (**Ho: $\sigma_1 = \sigma_2$**).
  - **Conclusion to test for equality of variances:** We conclude there is enought evidence of a difference in the variance of looks vs. personality score between males and females.
  - This results in using the row for Equal variances NOT assumed to find the t-test results including the test statistic, p-value, and confidence interval for the difference. (Outlined in **BLUE**)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence ... Lower	95% Confidence ... Upper
bmi	Equal variances assumed	10.491	.001	3.353	498	.001	1.63780245	.48847915	.67806843	2.59753647
	Equal variances not assumed			3.207	360.513	.001	1.63780245	.51073138	.63341547	2.64218943

- **SOFTWARE OUTPUT In SAS:**
  - The p-value for the test of equality of variances is reported as **0.0004 in the Pr > F column under equality of variances**.
  - So we reject the null hypothesis that the variances, or equivalently the standard deviations, are equal (**Ho: $\sigma_1 = \sigma_2$**).
  - **Conclusion to test for equality of variances:** We conclude there is enough evidence of a difference in the variance of looks vs. personality score between males and females.
  - This results in using the row for SATTERTHWAITE method where UNEQUAL variances are assumed to find the t-test results including the test statistic, p-value, and confidence interval for the difference. (Outlined in **BLUE**)

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		27.2689	26.7203	27.8175	4.8284	4.4705	5.2491
1		25.6311	24.7872	26.4750	6.0520	5.5114	6.7113
Diff (1-2)	Pooled	1.6378	0.6781	2.5975	5.3510	5.0384	5.7054
Diff (1-2)	Satterthwaite	1.6378	0.6334	2.6422			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	498	3.35	0.0009
Satterthwaite	Unequal	360.51	3.21	0.0015

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	199	299	1.57	0.0004

- **TEST STATISTIC for Two-Sample T-test:** In all of the results above, we determine that we will use the test which assumes the variances are UNEQUAL, and we find our **test statistic** of **t = 3.21.**

Step 3: Find the p-value of the test by using the test statistic as follows

- We will let the software find the p-value for us, and in this case, **the p-value is less than our significance level of 0.05.**
- This is found in **SPSS in the UNEQUAL variances assumed row under t-test in the SIG. (two-tailed) column given as 0.001** and in **SAS in the SATTERTHWAITE ROW under Pr > |t| column given as 0.0015.**
- This p-value means that it would be extremely rare to get data like that observed (or even more extreme) had the null hypothesis been true.
- More specifically, in our example, if there were no differences between females and males with respect to BMI, it would be almost highly unlikely (0.001 probability) to get data where the difference between the sample mean BMIs of males and females is 1.64 or more extreme.
- **Comment:** Note that the output tells us that the difference $\mu_1 - \mu_2$ is approximately 1.64. But more importantly, we want to know if this difference is statistically significant. To answer this, we use the fact that this difference is 3.21 standard errors above the null value.

Step 4: Conclusion

As usual a small p-value provides evidence against Ho. In our case our p-value is 0.001 (which is smaller than any level of significance that we will choose). The data therefore provide very strong evidence against Ho so we reject it.

- **Conclusion: The mean BMI of males differs from that of females. In other words, males and females differ with respect to BMI among heart attack patients.**

As a follow-up to this conclusion, we can construct a confidence interval for the difference between population means. In this case we will construct a confidence interval for $\mu_1 - \mu_2$ the population mean BMI for males minus the population mean BMI for females.

- Using statistical software, we find that the 95% confidence interval for $\mu_1 - \mu_2$ is roughly **(0.63, 2.64).**
- This is found in **SPSS in the UNEQUAL variances assumed row under 95% confidence interval columns** and in **SAS in the SATTERTHWAITE ROW under 95% CL MEAN column.**
- **Interpretation:**
  - **With 95% confidence that the population mean BMI for males is between 0.63 and 2.64 units larger than that of females.**
  - OR
  - **With 95% confidence that the population mean BMI for females is between 0.63 and 2.64 units smaller than that of males.**
- The confidence interval therefore quantifies the effect of the explanatory variable (gender) on the response (BMI). Notice that we cannot imply a causal effect of gender on BMI based upon this result alone as there could be many lurking variables, unaccounted for in this analysis, which might be partially or even completely responsible for this difference.

- **Note:** The confidence interval does not contain zero (both values are positive based upon how we chose our groups) and thus using the confidence interval we can reject the null hypothesis here.

Practical Significance:

- We should definitely ask ourselves if this is practically significant
- Is a true difference in population means as represented by our estimate from this data meaningful here? Is a difference in BMI of between 0.53 and 2.64 of interest?
- I will let you consider and answer for yourself.

SPSS Output for this example (Non-Parametric Output for Examples 1 and 2)

SAS Output and SAS Code (Includes Non-Parametric Test)
Note: In the SAS output the variable gender is not formatted, in this case Males = 0 and Females = 1.

Comments:

You might ask yourself: "Where do we use the test statistic?"

It is true that for all practical purposes all we have to do is check that the conditions which allow us to use the two-sample t-test are met, lift the p-value from the output, and draw our conclusions accordingly.

However, we feel that it is important to mention the test statistic for two reasons:

- The test statistic is what's behind the scenes; based on its null distribution and its value, the p-value is calculated.
- Apart from being the key for calculating the p-value, the test statistic is also itself a measure of the evidence stored in the data against Ho. As we mentioned, it measures (in standard errors) how different our data is from what is claimed in the null hypothesis.

Now try some more activities for yourself.

Did I Get This? Two-Sample T-test and Related Confidence Interval
(Non-Interactive Version – Spoiler Alert)

Non-Parametric Alternative: Wilcoxon Rank-Sum Test (Mann-Whitney U)

Learning Objectives

LO 5.1: For a data analysis situation involving two variables, determine the appropriate alternative (non-parametric) method when assumptions of our standard methods are not met.

We will look at one non-parametric test in the two-independent samples setting. More details will be discussed later (Details for Non-Parametric Alternatives).

- The **Wilcoxon rank-sum test (Mann-Whitney U test)** is a general test to compare two distributions in independent samples. It is a commonly used alternative to the two-sample t-test when the assumptions are not met.

k > 2 Independent Samples

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.35: For a data analysis situation involving two variables, choose the appropriate inferential method for examining the relationship between the variables and justify the choice.

Learning Objectives

LO 4.36: For a data analysis situation involving two variables, carry out the appropriate inferential method for examining relationships between the variables and draw the correct conclusions in context.

CO-5: Determine preferred methodological alternatives to commonly used statistical methods when assumptions are not met.

REVIEW: Unit 1 Case C-Q

Video

Video: 2 Independent Samples">k > 2 Independent Samples (21:15)

Related SAS Tutorials

- 7A (2:32) Numeric Summaries by Groups
- 7B (3:03) Side-By-Side Boxplots
- 7D (4:07) One Way ANOVA

Related SPSS Tutorials

- 7A (3:29) Numeric Summaries by Groups
- 7B (1:59) Side-By-Side Boxplots
- 7D (4:22) One Way ANOVA

Introduction

In this part, we continue to handle situations involving one categorical explanatory variable and one quantitative response variable, which is case C→Q.

Here is a summary of the tests we have covered for the case where k = 2. Methods in **BOLD** are our main focus in this unit.

So far we have discussed the two samples and matched pairs designs, in which the categorical explanatory variable is two-valued. As we saw, in these cases, examining the relationship between the explanatory and the response variables amounts to comparing the mean of the response variable (Y) in two populations, which are defined by the two values of the explanatory variable (X). The difference between the two samples and matched pairs designs is that in the former, the two samples are independent, and in the latter, the samples are dependent.

Independent Samples (More Emphasis)	Dependent Samples (Less Emphasis)
Standard Tests • **Two Sample T-Test Assuming Equal Variances** • **Two Sample T-Test Assuming Unequal Variances** Non-Parametric Test • Mann-Whitney U (or Wilcoxon Rank-Sum) Test	**Standard Test** • **Paired T-Test** Non-Parametric Tests • Sign Test • Wilcoxon Signed-Rank Test

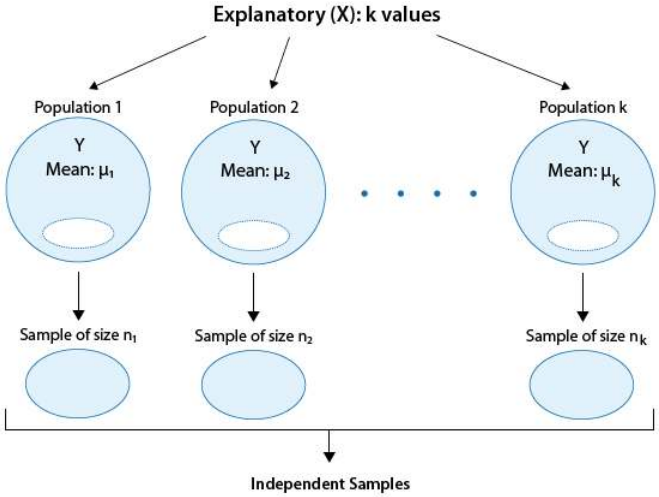
We now move on to the case where k > 2 when we have independent samples. Here is a summary of the tests we will learn for the case where k > 2. **Notice we will not cover the dependent samples case in this course.**

Independent Samples (Only Emphasis)	Dependent Samples (Not Discussed)

Standard Tests • **One-way ANOVA (Analysis of Variance)** Non-Parametric Test • Kruskal–Wallis One-way ANOVA	*Standard Test* • *Repeated Measures ANOVA (or similar)*

Here, as in the two-valued case, making inferences about the relationship between the explanatory (X) and the response (Y) variables amounts to comparing the means of the response variable in the populations defined by the values of the explanatory variable, where the number of means we are comparing depends, of course, on the number of values of X.

Unlike the two-valued case, where we looked at two sub-cases (1) when the samples are independent (two samples design) and (2) when the samples are dependent (matched pairs design, here, we are just going to discuss the case where the samples are independent. In other words, we are just going to extend the two samples design to more than two independent samples.



The inferential method for comparing more than two means that we will introduce in this part is called **ANalysis Of VAriance** (abbreviated as **ANOVA**), and the test associated with this method is called the ANOVA F-test.

In most software, the data need to be arranged so that each row contains one observation with one variable recording X and another variable recording Y for each observation.

Comparing Two or More Means – The ANOVA F-test

Learning Objectives

LO 4.38: In a given context, determine the appropriate standard method for comparing groups and provide the correct conclusions given the appropriate software output.

Learning Objectives

LO 4.39: In a given context, set up the appropriate null and alternative hypotheses for comparing groups.

As we mentioned earlier, the test that we will present is called the ANOVA F-test, and as you'll see, this test is different in two ways from all the tests we have presented so far:

- Unlike the previous tests, where we had three possible alternative hypotheses to choose from (depending on the context of the problem), in the ANOVA F-test there is only one alternative, which actually makes life simpler.
- The test statistic will **not** have the same structure as the test statistics we've seen so far. In other words, it will **not** have the form:

$$\text{test statistic} = \frac{\text{estimator - null value}}{\text{standard error of estimator}}$$

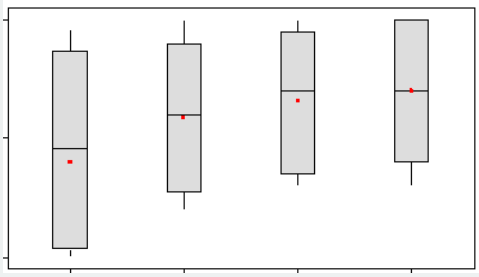
but a different structure that captures the essence of the F-test, and clarifies where the name “analysis of variance” is coming from.

What is the idea behind comparing more than two means?

The question we need to answer is: Are the differences among the sample means due to true differences among the μ’s (alternative hypothesis), or merely due to sampling variability or random chance (null hypothesis)?

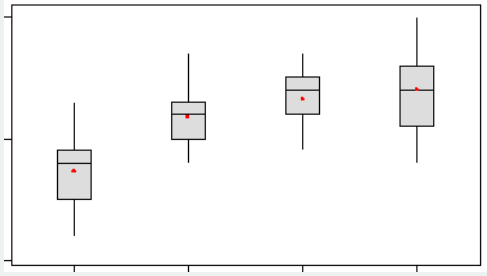
Here are two sets of boxplots representing two possible scenarios:

Scenario #1



- Because of the large amount of spread within the groups, this data shows boxplots with plenty of overlap.
- One could imagine the data arising from 4 random samples taken from 4 populations, all having the same mean of about 11 or 12.
- The first group of values may have been a bit on the low side, and the other three a bit on the high side, but such differences could conceivably have come about by chance.
- This would be the case if the null hypothesis, claiming equal population means, were true.

Scenario #2



- Because of the small amount of spread within the groups, this data shows boxplots with very little overlap.
- It would be very hard to believe that we are sampling from four groups that have equal population means.
- This would be the case if the null hypothesis, claiming equal population means, were false.

Thus, in the language of hypothesis tests, we would say that if the data were configured as they are in scenario 1, we would not reject the null hypothesis that population means were equal for the k groups.

If the data were configured as they are in scenario 2, we would reject the null hypothesis, and we would conclude that not all population means are the same for the k groups.

Let’s summarize what we learned from this.

- The question we need to answer is: Are the differences among the sample means due to true differences among the μ’s (alternative hypothesis), or merely due to sampling variability (null hypothesis)?

In order to answer this question using data, we need to look at the variation among the sample means, but this alone is not enough.

We need to look at the variation among the sample means relative to the variation within the groups. In other words, we need to look at the quantity:

$$\frac{\text{VARIATION AMONG SAMPLE MEANS}}{\text{VARIATION WITHIN GROUPS}}$$

which measures to what extent the difference among the sample means for our groups dominates over the usual variation within sampled groups (which reflects differences in individuals that are typical in random samples).

When the variation within groups is large (like in scenario 1), the variation (differences) among the sample means may become negligible resulting in data which provide very little evidence against Ho. When the variation within groups is small (like in scenario 2), the variation among the sample means dominates over it, and the data have stronger evidence against Ho.

It has a different structure from all the test statistics we've looked at so far, but it is similar in that it is still a measure of the evidence against H_0. The larger F is (which happens when the denominator, the variation within groups, is small relative to the numerator, the variation among the sample means), the more evidence we have against H_0.

Looking at this ratio of variations is the idea behind the comparing more than two means; hence the name analysis of variance (ANOVA).

Now test your understanding of this idea.

Learn By Doing: Idea of One-Way ANOVA
(Non-Interactive Version – Spoiler Alert)

Comments

- The focus here is for you to understand the idea behind this test statistic, so we do not go into detail about how the two variations are measured. We instead rely on software output to obtain the F-statistic.
- This test is called the ANOVA F-test.
  - So far, we have explained the ANOVA part of the name.
  - Based on the previous tests we introduced, it should not be surprising that the "F-test" part comes from the fact that the null distribution of the test statistic, under which the p-values are calculated, is called an F-distribution.
  - We will say very little about the F-distribution in this course, which will essentially be limited to this comment and the next one.
- It is fairly straightforward to decide if a z-statistic is large. Even without tables, we should realize by now that a z-statistic of 0.8 is not especially large, whereas a z-statistic of 2.5 is large.
  - In the case of the t-statistic, it is less straightforward, because there is a different t-distribution for every sample size n (and degrees of freedom $n - 1$).
  - However, the fact that a t-distribution with a large number of degrees of freedom is very close to the z (standard normal) distribution can help to assess the magnitude of the t-test statistic.
  - When the size of the F-statistic must be assessed, the task is even more complicated, because there is a different F-distribution for every combination of the number of groups we are comparing and the total sample size.
  - We will nevertheless say that for most situations, an F-statistic greater than 4 would be considered rather large, but tables or software are needed to get a truly accurate assessment.

Steps for One-Way ANOVA

Here is a **full statement of the process for the ANOVA F-Test:**

Step 1: State the hypotheses

The null hypothesis claims that there is no relationship between X and Y. Since the relationship is examined by comparing the means of Y in the populations defined by the values of X (μ_1, μ_2, …, μ_k), no relationship would mean that all the means are equal.

Therefore the **null hypothesis of the F-test is:**

- **Ho:** $\mu_1 = \mu_2 = \ldots = \mu_k$. (There is no relationship between X and Y.)

As we mentioned earlier, here we have just **one alternative hypothesis**, which claims that there **is** a relationship between X and Y. In terms of the means $\mu_1, \mu_2, \ldots, \mu_k$, it simply says the opposite of the null hypothesis, that not all the means are equal, and we simply write:

- **Ha:** not all μ's are equal. (There is a relationship between X and Y.)

Learn By Doing: One-Way ANOVA – STEP 1
(Non-Interactive Version – Spoiler Alert)

Comments:

- The alternative of the ANOVA F-test simply states that not all of the means are equal, and is not specific about the way in which they are different.
- Another way to phrase the alternative is
  - Ha: at least two μ's are different
- **Warning:** It is incorrect to say that the alternative is $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$. This statement is MUCH stronger than our alternative hypothesis and says ALL means are different from ALL other mean
- Note that there are many ways for μ1, μ2, μ3, μ4 not to be all equal, and μ1 ≠ μ2 ≠ μ3 ≠ μ4 is just one of them. Another way could be μ1 = μ2 = μ3 ≠ μ4 or μ1 = μ2 ≠ μ3 = μ4. The alternative of the ANOVA F-test simply states that not all of the means are equal, and is not specific about the way in which they are different.

Step 2: Obtain data, check conditions, and summarize data

The ANOVA F-test can be safely used as long as the following conditions are met:

- The samples drawn from each of the populations we're comparing are independent.
- We are in one of the following two scenarios:

  (i) Each of the populations are normal, or more specifically, the distribution of the response Y in each population is normal, and the samples are random (or at least can be considered as such). In practice, checking normality in the populations is done by looking at each of the samples using a histogram and checking whether there are any signs that the populations are not normal. Such signs could be extreme skewness and/or extreme outliers.

  (ii) The populations are known or discovered not to be normal, but the sample size of each of the random samples is large enough (we can use the rule of thumb that a sample size greater than 30 is considered large enough).

- The populations all have the same standard deviation.

  Can check this condition using the rule of thumb that the ratio between the largest sample standard deviation and the smallest is less than 2. If that is the case, this condition is considered to be satisfied.

  Can check this condition using a formal test similar to that used in the two-sample t-test although we will not cover any formal tests.

Learn By Doing: One-Way ANOVA – STEP 2
(Non-Interactive Version – Spoiler Alert)

Test Statistic

- If our conditions are satisfied we have the test statistic.

$$F = \frac{\text{VARIATION AMONG SAMPLE MEANS}}{\text{VARIATION WITHIN GROUPS}}$$

- The statistic follows an F-distribution with k-1 numerator degrees of freedom and n-k denominator degrees of freedom.
- Where n is the total (combined) sample size and k is the number of groups being compared.
- We will rely on software to calculate the test statistic and p-value for us.

Step 3: Find the p-value of the test by using the test statistic as follows

- The p-value of the ANOVA F-test is the probability of getting an F statistic as large as we obtained (or even larger), had Ho been true (all k population means are equal).
- In other words, it tells us how surprising it is to find data like those observed, assuming that there is no difference among the population means $\mu_1, \mu_2, \ldots, \mu_k$.

Step 4: Conclusion

As usual, we base our conclusion on the p-value.

- A **small p-value** tells us that our data contain a lot of evidence against Ho. More specifically, a small p-value tells us that the differences between the sample means are statistically significant (unlikely to have happened by chance), and therefore **we reject Ho.**
  - **Conclusion:** There is enough evidence that the categorical explanatory variable is related to (or associated with) the quantitative response variable. More specifically, there is enough evidence that there are differences between at least two of the population means (there are some differences in the population means).
- **If the p-value is not small**, we do not have enough statistical evidence to reject Ho.
  - **Conclusion:** There is NOT enough evidence that the categorical explanatory variable is related to (or associated with) the quantitative response variable. More specifically, there is NOT enough evidence that there are differences between at least two of the population means.
- A significance level (cut-off probability) of 0.05 can help determine what is considered a small p-value.

Final Comment

Note that when we reject Ho in the ANOVA F-test, all we can conclude is that

- not all the means are equal, or
- there are some differences between the means, or
- the response Y is related to explanatory X.

However, the ANOVA F-test does not provide any immediate insight into why Ho was rejected, or in other words, it does not tell us in what way the population means of the groups are different. As an exploratory (or visual) aid to get that insight, we may take a look at the confidence intervals for group population means. More specifically, we can look at which of the confidence intervals overlap and which do not.

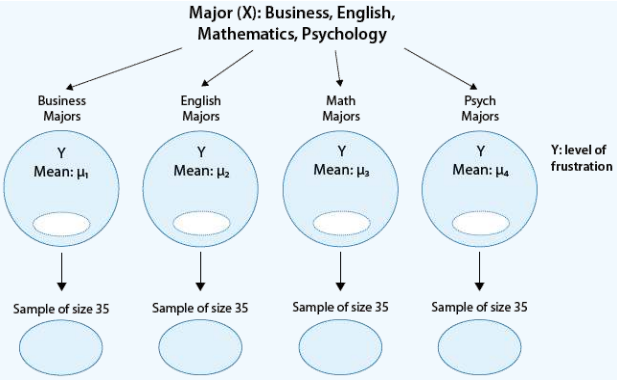
Multiple Comparisons:

- When we compare standard 95% confidence intervals in this way, we have an increased chance of making a type I error as each interval has a 5% error individually.
- There are many multiple comparison procedures all of which propose alternative methods for determining which pairs of means are different.
- We will look at a few of these in the software just to show you a little about this topic but we will not cover this officially in this course.
- The goal is to provide an overall type I error rate no larger than 5% for all comparisons made.

Now let's look at some examples using real data.

EXAMPLE: Is "academic frustration" related to major?

A college dean believes that students with different majors may experience different levels of academic frustration. Random samples of size 35 of Business, English, Mathematics, and Psychology majors are asked to rate their level of academic frustration on a scale of 1 (lowest) to 20 (highest).



The figure highlights what we have already mentioned: examining the relationship between major (X) and frustration level (Y) amounts to comparing the mean frustration levels among the four majors defined by X. Also, the figure reminds us that we are dealing with a case where the samples are independent.

Step 1: State the hypotheses

The correct hypotheses are:

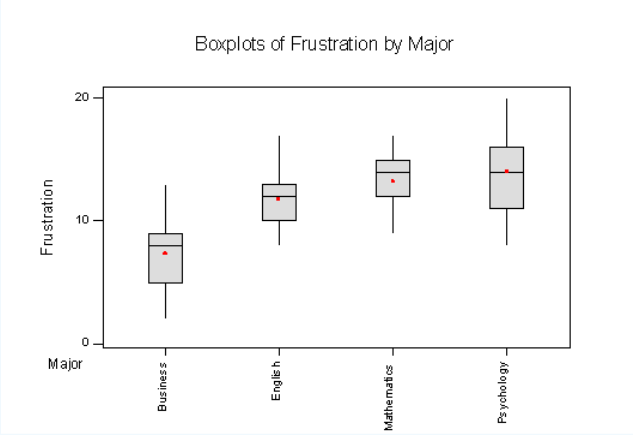
- **Ho:** $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
  (There is **NO** relationship between major and academic frustration level.)
- **Ha:** not all μ's are equal.
  (There **IS** a relationship between major and academic frustration level.)

Step 2: Obtain data, check conditions, and summarize data

Data: SPSS format, SAS format, Excel format, CSV format

In our example all the conditions are satisfied:

- All the samples were chosen randomly, and are therefore independent.
- The sample sizes are large enough (n = 35) that we really don't have to worry about the normality; however, let's look at the data using side-by-side boxplots, just to get a sense of it:



- The data suggest that the frustration level of the business students is generally lower than students from the other three majors. The ANOVA F-test will tell us whether these differences are significant.

The rule of thumb is satisfied since 3.082 / 2.088 < 2. We will look at the formal test in the software.

Summary statistics:

Column	n	Mean	Std. Err.	Std. Dev.	Min	Q1	Median	Q3	Max
Business	35	7.3142858	0.48984894	2.8979855	2	5	8	9	13
English	35	11.771428	0.35286513	2.0875783	8	10	12	13	17
Mathematics	35	13.2	0.3639189	2.1529734	9	12	14	15	17
Psychology	35	14.028571	0.52096504	3.0820706	8	11	14	16	20

Test statistic: (Minitab output)

One-way ANOVA: Frustration Score versus Major

```
Source   DF       SS      MS      F      P
Major     3   939.85  313.28  46.60  0.000
Error   136   914.29    6.72
Total   139  1854.14

S = 2.593   R-Sq = 50.69%   R-Sq(adj) = 49.60%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level         N    Mean  StDev  ----+---------+---------+---------+-----
Business     35   7.314  2.898  (--*---)
English      35  11.771  2.088                   (--*---)
Mathematics  35  13.200  2.153                       (---*--)
Psychology   35  14.029  3.082                           (--*---)
                              ----+---------+---------+---------+-----
                                7.5      10.0      12.5      15.0

Pooled StDev = 2.593
```

- The parts of the output that we will focus on here have been highlighted. In particular, note that the **F-statistic is 46.60**, which is very large, indicating that the data provide a lot of evidence against Ho (we can also see that the p-value is so small that it is reported to be 0, which supports that conclusion as well).

Step 3: Find the p-value of the test by using the test statistic as follows

- As we already noticed before, the p-value in our example is so small that it is reported to be 0.000, telling us that it would be next to impossible to get data like those observed had the mean frustration level of the four majors been the same (as the null hypothesis claims).

Step 4: Conclusion

- In our example, **the p-value is extremely small – close to 0 –** indicating that our data provide extremely strong evidence to reject Ho.
- **Conclusion:** There is enough evidence that the population mean frustration level of the four majors are not all the same, or in other words, that majors do have an effect on students' academic frustration levels at the school where the test was conducted.

As a follow-up, we can construct confidence intervals (or conduct multiple comparisons as we will do in the software). This allows us to understand better which population means are likely to be different.

One-way ANOVA: Frustration Score versus Major

```
Source   DF       SS      MS      F      P
Major     3   939.85  313.28  46.60  0.000
Error   136   914.29    6.72
Total   139  1854.14

S = 2.593   R-Sq = 50.69%   R-Sq(adj) = 49.60%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level         N    Mean  StDev  ----+---------+---------+---------+-----
Business     35   7.314  2.898  (--*---)
English      35  11.771  2.088                   (--*---)
Mathematics  35  13.200  2.153                       (---*--)
Psychology   35  14.029  3.082                           (--*---)
                              ----+---------+---------+---------+-----
                                7.5      10.0      12.5      15.0

Pooled StDev = 2.593
```

In this case, the business majors are clearly lower on the frustration scale than other majors. It is also possible that English majors are lower than psychology majors based upon the individual 95% confidence intervals in each group.

SPSS Output

SAS Output and SAS Code (Includes Non-Parametric Test)

Here is another example

EXAMPLE: Reading Level in Adversting

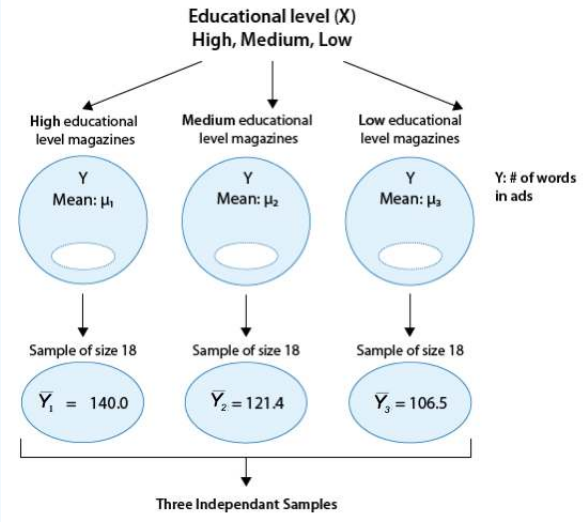
Do advertisers alter the reading level of their ads based on the target audience of the magazine they advertise in?

In 1981, a study of magazine advertisements was conducted (F.K. Shuptrine and D.D. McVicker, "Readability Levels of Magazine Ads," Journal of Advertising Research, 21:5, October 1981). Researchers selected random samples of advertisements from each of three groups of magazines:

- Group 1—highest educational level magazines (such as Scientific American, Fortune, The New Yorker)
- Group 2—middle educational level magazines (such as Sports Illustrated, Newsweek, People)
- Group 3—lowest educational level magazines (such as National Enquirer, Grit, True Confessions)

The measure that the researchers used to assess the level of the ads was the number of words in the ad. 18 ads were randomly selected from each of the magazine groups, and the number of words per ad were recorded.

The following figure summarizes this problem:



Our question of interest is whether the number of words in ads (Y) is related to the educational level of the magazine (X). To answer this question, we need to compare μ_1, μ_2, and μ_3, the mean number of words in ads of the three magazine groups. Note in the figure that the sample means are provided. It seems that what the data suggest makes sense; the magazines in group 1 have the largest number of words per ad (on average) followed by group 2, and then group 3.

The question is whether these differences between the sample means are significant. In other words, are the differences among the observed sample means due to true differences among the μ's or merely due to sampling variability? To answer this question, we need to carry out the ANOVA F-test.

Step 1: Stating the hypotheses.

We are testing:

- **Ho:** $\mu_1 = \mu_2 = \mu_3$.
  (There is **NO** relationship between educational level and number of words in ads.)
- **Ha:** not all μ's are equal.
  (There **IS** a relationship between educational level and number of words in ads.)

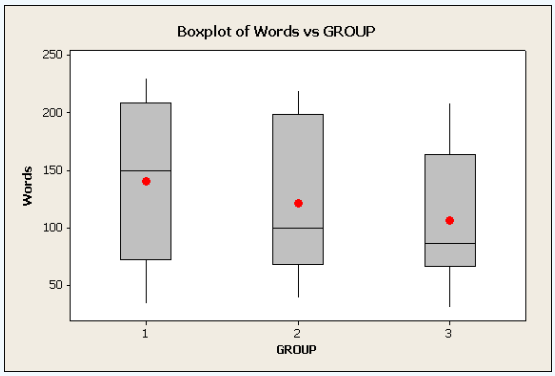
Conceptually, the null hypothesis claims that the number of words in ads is not related to the educational level of the magazine, and the alternative hypothesis claims that there is a relationship.

Step 2: Checking conditions and summarizing the data.

- (i) The ads were selected at random from each magazine group, so the three samples are independent.

In order to check the next two conditions, we'll need to look at the data (condition ii), and calculate the sample standard deviations of the three samples (condition iii).

- Here are the side-by-side boxplots of the data:



- And the standard deviations:
  - Group 1 StDev: 74.0
  - Group 2 StDev: 64.3
  - Group 3 StDev: 57.6

Using the above, we can address conditions (ii) and (iii)

- (ii) The graph does not display any alarming violations of the normality assumption. It seems like there is some skewness in groups 2 and 3, but not extremely so, and there are no outliers in the data.
- (iii) We can assume that the equal standard deviation assumption is met since the rule of thumb is satisfied: the largest sample standard deviation of the three is 74 (group 1), the smallest one is 57.6 (group 3), and 74/57.6 < 2.

Before we move on, let's look again at the graph. It is easy to see the trend of the sample means (indicated by red circles).

However, there is so much variation within each of the groups that there is almost a complete overlap between the three boxplots, and the differences between the means are over-shadowed and seem like something that could have happened just by chance.

Let's move on and see whether the ANOVA F-test will support this observation.

- **Test Statistic:** Using statistical software to conduct the ANOVA F-test, we find that the **F statistic is 1.18**, which is not very large. We also find that the p-value is 0.317.

Step 3. Finding the p-value.

- **The p-value is 0.317**, which tells us that getting data like those observed is not very surprising assuming that there are no differences between the three magazine groups with respect to the mean number of words in ads (which is what H_0 claims).
- In other words, the large p-value tells us that it is quite reasonable that the differences between the observed sample means could have happened just by chance (i.e., due to sampling variability) and not because of true differences between the means.

Step 4: Making conclusions in context.

- The large p-value indicates that the results are not statistically significant, and that we cannot reject H_0.
- **Conclusion:** The study does not provide evidence that the mean number of words in ads is related to the educational level of the magazine. In other words, the study does not provide evidence that advertisers alter the reading level of their ads (as measured by the number of words) based on the educational level of the target audience of the magazine.

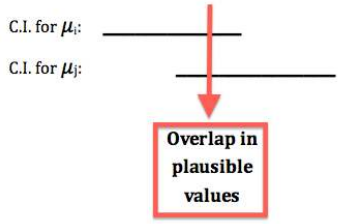
Now try one for yourself.

Learn By Doing: One-Way ANOVA – Flicker Frequency
(Non-Interactive Version – Spoiler Alert)

Confidence Intervals

The ANOVA F-test does not provide any insight into why H_0 was rejected; it does not tell us in what way μ1,μ2,μ3…,μk are not all equal. We would like to know which pairs of 's are not equal. As an exploratory (or visual) aid to get that insight, we may take a look at the confidence intervals for group population meansμ1,μ2,μ3…,μk that appears in the output. More specifically, we should look at the position of the confidence intervals and overlap/no overlap between them.

* If the confidence interval for, say,μi overlaps with the confidence interval for μj , then μi and μj share some plausible values, which means that based on the data we have no evidence that these two 's are different.

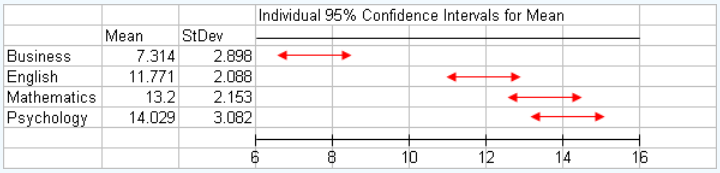


* If the confidence interval for μi does not overlap with the confidence interval for μj , then μi and μj do not share plausible values, which means that the data suggest that these two 's are different.

Furthermore, if like in the figure above the confidence interval (set of plausible values) for μi lies entirely below the confidence interval (set of plausible values) for μj, then the data suggest that μi is smaller than μj.

EXAMPLE

Consider our first example on the level of academic frustration.



Based on the small p-value, we rejected H_0 and concluded that not all four frustration level means are equal, or in other words that frustration level is related to the student's major. To get more insight into that relationship, we can look at the confidence intervals above (marked in red). The top confidence interval is the set of plausible values for μ_1, the mean frustration level of business students. The confidence interval below it is the set of plausible values for μ_2, the mean frustration level of English students, etc.

What we see is that the business confidence interval is way below the other three (it doesn't overlap with any of them). The math confidence interval overlaps with both the English and the psychology confidence intervals; however, there is no overlap between the English and psychology confidence intervals.

This gives us the impression that the mean frustration level of business students is lower than the mean in the other three majors. Within the other three majors, we get the impression that the mean frustration of math students may not differ much

from the mean of both English and psychology students, however the mean frustration of English students may be lower than the mean of psychology students.

Note that this is only an exploratory/visual way of getting an impression of why H_o was rejected, not a formal one. There is a formal way of doing it that is called "multiple comparisons," which is beyond the scope of this course. An extension to this course will include this topic in the future.

Non-Parametric Alternative: Kruskal-Wallis Test

Learning Objectives

LO 5.1: For a data analysis situation involving two variables, determine the appropriate alternative (non-parametric) method when assumptions of our standard methods are not met.

We will look at one non-parametric test in the k > 2 independent sample setting. We will cover more details later (Details for Non-Parametric Alternatives).

The Kruskal-Wallis test is a general test to compare multiple distributions in independent samples and is a common alternative to the one-way ANOVA.

Details for Non-Parametric Alternatives in Case C-Q

Learn By Doing: Supplemental Examples and Exercises for Unit 4B
(Non-interactive Version)

Caution

As we mentioned at the end of the Introduction to Unit 4B, **we will focus only on two-sided tests** for the remainder of this course. One-sided tests are often possible but rarely used in clinical research.

CO-5: Determine preferred methodological alternatives to commonly used statistical methods when assumptions are not met.

Video

Video: Details for Non-Parametric Alternatives (17:38)

Related SAS Tutorials

- 7E – (4:34) Non Parametric Tests for independent samples (k= 2 and k > 2)
- 8C – (5:20) Paired T-Test and Non Parametric Tests for dependent samples

Related SPSS Tutorials

- 7E – (3:57) Non Parametric Tests for independent samples (k= 2 and k > 2)
- 8D – (3:32) Non Parametric (Paired) for dependent samples

We mentioned some non-parametric alternatives to the paired t-test, two-sample t-test for independent samples, and the one-way ANOVA.

Here we provide more details and resources for these tests for those of you who wish to conduct them in practice.

Non-Parametric Tests

The statistical tests we have previously discussed require assumptions about the distribution in the population or about the requirements to use a certain approximation as the sampling distribution. These methods are called **parametric**.

When these assumptions are not valid, alternative methods often exist to test similar hypotheses. Tests which require only minimal distributional assumptions, if any, are called **non-parametric** or **distribution-free** tests.

In some cases, these tests may be called **exact tests** due to the fact that their methods of calculating p-values or confidence intervals require no mathematical approximation (a foundation of many statistical methods).

However, note that when the assumptions are precisely satisfied, some "parametric" tests can also be considered "exact."

Case CQ – Matched Pairs

We will look at two non-parametric tests in the paired sample setting.

The Sign Test

The sign test is a very general test used to compare paired samples. It can be used instead of the Paired T-test if the assumptions are not met although the next test we discuss is likely a better option in that case as we will see. However, the sign test does have some advantages and is worth understanding.

- The idea behind the test is to find the **sign of the differences (positive or negative)** and use this information to determine if the medians between the two groups are the same.
- If the two paired measurements came from the populations with equal medians, we would expect half of the differences to be positive and half to be negative. Thus the sampling distribution of our statistic is simply a binomial with p = 0.5.

Outline of Procedure for the SIGN TEST

- **Step 1: State the hypotheses**

The hypotheses are:

Ho: the medians are equal

Ha: the medians are not equal (one-sided tests are possible)

- **Step 2: Obtain data, check conditions, and summarize data**

We require a random sample (or at least can be considered random in context).

The sign test can be used for any data for which the sign of the difference can be obtained. Thus, it can be used for:

quantitative measures (continuous or discrete)
Examples: Systolic Blood Pressure, Number of Drinks

(categorical) ordinal measures
Examples: Rating scales, Letter Grades

(categorical) binary measures where we can only tell whether one pair is "larger" or "smaller" compared to the other pair
Examples: Is the left arm more or less sunburned than the right arm?, Was there an improvement in pain after treatment?

For this reason, this test is very widely applicable!

The data are summarized by a test statistic which counts the number of positive (or negative) differences. Any ties (zero differences) are discarded.

- **Step 3: Find the p-value of the test by using the test statistic as follows**

The p-values are calculated using the binomial distribution (or a normal approximation for large samples). We will rely on software to obtain the p-value for this test.

- **Step 4: Conclusion**

The decision is made in the same manner as other tests.

We can word our conclusion in terms of the medians in the two populations or in terms of the relationship between the categorical explanatory variable (X) and the response variable (Y).

OPTIONAL: For more details visit The Sign Test in Penn State's online content for STAT 415.

The Wilcoxon Signed-Rank Test:

The Wilcoxon signed-rank Test is a general test to compare distributions in paired samples. This test is usually the preferred alternative to the Paired t-test when the assumptions are not satisfied.

The idea behind the test is to determine if the two populations seem to be the same or different based upon the ranks of the absolute differences (instead of the magnitude of the differences). Ranking procedures are commonly used in non-parametric methods as this moderates the effect of any outliers.

We have one assumption for this test. We assume the distribution of the differences is symmetric.

Under this assumption, if the two paired measurements came from the populations with equal means/medians, we would expect the two sets of ranks (those for positive differences and those for negative differences) to be distributed similarly. If there is a large difference here, this gives evidence of a true difference.

Outline of Procedure for the Wilcoxon Signed-Rank Test

- **Step 1: State the hypotheses**

The hypotheses are:

Ho: the means/medians are equal

Ha: the means/medians are not equal (one-sided tests are possible)

- **Step 2: Obtain data, check conditions, and summarize data**

We have a random sample and we assume the distribution of the differences is symmetric so we should check to be sure that there is no clear skewness to the distribution of the differences.

The Wilcoxon signed-Rank test can be used for quantitative or ordinal data (but not binary as for the sign test).

The data are summarized by a test statistic which counts the sum of the positive (or negative) ranks. Any zero differences are discarded.

To rank the pairs, we find the differences (much as we did in the paired t-test), take the absolute value of these differences and rank the pairs from 1 = smallest non-zero difference to m = largest non-zero difference, where m = number of non-zero pairs.

Then we determine which ranks came from positive (or negative) differences and find the sum of these ranks.

You will not be conducting this test by hand. We simply wish to explain some of the logic behind the scenes for these tests.

- **Step 3: Find the p-value of the test by using the test statistic as follows**

The p-values are calculated using a distribution specific to this test. We will rely on software to obtain the p-value for this test.

- **Step 4: Conclusion**

The decision is made in the same manner as other tests. We can word our conclusion in terms of the means or medians in the two populations or in terms of the existence or non-existence of a relationship between the categorical explanatory variable (X) and the response variable (Y).

OPTIONAL: For more details on these tests visit The Wilcoxon Signed Rank Test in Penn State's online content for STAT 415.

Comments:

- The **sign test tends to have much lower power than the paired t-test or the Wilcoxon signed-Rank test**. In other words, the sign test has less chance of being able to detect a true difference than the other tests. It is, however, applicable in the case where we only know "better" or "worse" for each pair, where the other two methods are not.

- The **Wilcoxon signed-rank test is comparable to the paired t-test in power and can even perform better than the paired t-test under certain conditions**. In particular, this can occur when there are a few very large outliers as these outliers can greatly affect our estimate of the standard error in the paired t-test since it is based upon the sample standard deviation which is highly affected by such outliers.
- **Both the sign Test and the Wilcoxon signed-rank test can also be used for one sample**. In that case, you must specify the null value and calculate differences between the observed value and the null value (instead of the difference between two pairs).

Case CQ – Two Independent Samples – Wilcoxon Rank-Sum Test (Mann-Whitney U):

We will look at one non-parametric test in the two-independent samples setting.

The **Wilcoxon rank-sum test (Mann-Whitney U test)** is a general test to compare two distributions in independent samples. It is a commonly used alternative to the two-sample t-test when the assumptions are not met.

The idea behind the test is to determine if the two populations seem to be the same or different based upon the ranks of the values instead of the magnitude. Ranking procedures are commonly used in non-parametric methods as this moderates the effect of any outliers.

There are many ways to formulate this test. For our purposes, we will assume the quantitative variable (Y) is a continuous random variable (or can be treated as continuous, such as for very large counts) and that we are interested in testing whether there is a "shift" in the distribution. In other words, we assume that the distribution is the same except that in one group the distribution is higher (or lower) than in the other.

- **Step 1: State the hypotheses**

We assume the distributions of the two populations are the same except for a horizontal shift in location.

The hypotheses are:

Ho: the medians are equal

Ha: the medians are not equal (one-sided tests are possible)

- **Step 2: Obtain data, check conditions, and summarize data**

(i) We have two independent random samples. All observations in each sample must be independent of all other observations.

(ii) The version of the Wilcoxon rank-sum test (Mann-Whitney U test) we are using assumes a that the quantitative response variable is a continuous random variable.

(iii) We assume there is only a location shift so we should check that the two distributions are similar except possibly for their locations.

(iv) The data are summarized by a test statistic which counts the sum of the sample 1 (or sample 2) ranks.

To rank the observations, we combine all observations in both samples and rank from smallest to largest.

Then we determine which ranks came from sample 1 (or sample 2) and find the sum of these ranks.

You will not be conducting this test by hand. We simply wish to explain some of the logic behind the scenes for these tests.

- **Step 3: Find the p-value of the test by using the test statistic as follows**

The p-values are calculated using a distribution specific to this test. We will rely on software to obtain the p-value for this test.

- **Step 4: Conclusion**

The decision is made in the same manner as other tests. We can word our conclusion in terms of the medians in the two populations or in terms of the existence or non-existence of a relationship between the categorical explanatory variable (X) and the response variable (Y).

OPTIONAL: For more details on this test visit The Wilcoxon Rank-Sum Test from Boston University School of Public Health

Case CQ – K > 2 – The Kruskal-Wallis Test

We will look at one non-parametric test in the k > 2 independent sample setting.

The Kruskal-Wallis test is a general test to compare multiple distributions in independent samples.

The idea behind the test is to determine if the k populations seem to be the same or different based upon the ranks of the values instead of the magnitude. Ranking procedures are commonly used in non-parametric methods as this moderates the effect of any outliers.

The test assumes identically-shaped and scaled distributions for each group, except for any difference in medians.

Step 1: State the hypotheses The hypotheses are:

- **Ho:** the medians of all groups are equal
- **Ha:** the medians are not all equal

Step 2: Obtain data, check conditions, and summarize data

(i) We have independent random samples from our k populations. All observations in each sample must be independent of all other observations.

(ii) We have an ordinal, discrete, or continuous response variable Y.

(iii) We assume there is only a location shift so we should check that the distributions are similar except possibly for their locations.

(iv) The data are summarized by a test statistic which involves the ranks of observations in each group.

To rank the observations, we combine all observations in all samples and rank from smallest to largest.

Then we determine which ranks came from which sample and use these to obtain the test statistic.

Step 3: Find the p-value of the test by using the test statistic as follows

The p-values are calculated using a distribution specific to this test. We will rely on software to obtain the p-value for this test.

Step 4: Conclusion

The decision is made in the same manner as other tests. We can word our conclusion in terms of the medians in the k populations or in terms of the existence or non-existence of a relationship between the categorical explanatory variable (X) and the response variable (Y).

OPTIONAL: For more details on this test visit The Kruskal-Wallis Test from Boston University School of Public Health

Let's Summarize

- We have presented the basic idea for the non-parameteric alternatives for Case C-Q
  - The sign test and the Wilcoxon signed-rank test are possible alternatives to the paired t-test in the case of two dependent samples.
  - The Wilcoxon rank-sum test (also known as the Mann-Whitney U test) is a possible alternative to the two-sample t-test in the case of two independent samples.
  - The Kruskal-Wallis test is a possible alternative to the one-way ANOVA in the case of more than two independent samples.
- In this course, we simply want you to be aware of which non-parameteric alternatives are commonly used to address issues with the assumptions.
- We are not asking you to conduct these tests but we do still provide information for those interested in being able to conduct these tests in practice.

Wrap-Up (Case C-Q)

> **Learn By Doing:** Supplemental Examples and Exercises for Unit 4B
> (Non-interactive Version)

Caution

As we mentioned at the end of the Introduction to Unit 4B, **we will focus only on two-sided tests** for the remainder of this course. One-sided tests are often possible but rarely used in clinical research.

We are now done with case C→Q.

- We learned that this case is further classified into sub-cases, depending on the number of groups that we are comparing (i.e., the number of categories that the explanatory variable has), and the design of the study (independent vs. dependent samples).
- For each of the three sub-cases that we covered, we learned the appropriate inferential method, and emphasized the idea behind the method, the conditions under which it can be safely used, how to carry it out using software, and the interpretation of the results.
- We also learned which non-parametric tests are applicable and under what circumstances they might be used instead of the standard methods.

The following table summarizes when each of the three standard tests, covered in this module, are used:

Sub-Case of C→Q	Circumstances When Used

Paired t-test
(special case of the one sample t-test)
	• Categorical explanatory variable with two categories
	• Comparing the two population means, when the samples are dependent on each other or "matched pairs."
	• Samples are dependent in the sense that every observation in one sample is linked to an observation in another sample. Examples of dependent samples include:
	  ○ same subjects measured twice
	  ○ Twins

Two-Sample t-test	• Categorical explanatory variable with two categories • Comparing two population means based on two independent samples • Either normal populations or large sample size
ANOVA	• Categorical explanatory variable with more than two categories • Comparing more than two population means based on independent samples

The following summary discusses each of the above named sub-cases of C→Q within the context of the hypothesis testing process.

Step 1: Stating the null and alternative hypotheses (H_0 and H_a)

- Although the one-sided alternatives are provided here where possible, remember that we will focus only on two-sided tests supplemented by confidence intervals for methods in Unit 4B.

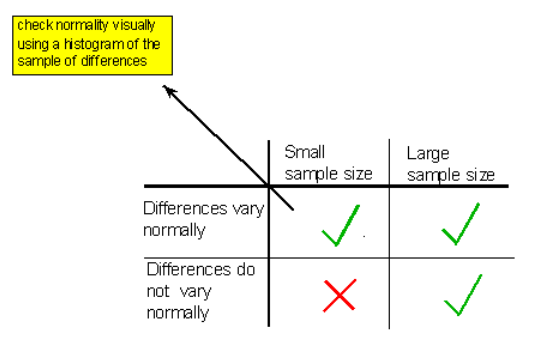

In a Two-Sample t-test, the hypotheses are: H_0: μ_1 - μ_2 = 0 (or H_0: μ_1 = μ_2), and one of: * H_a: μ_1 - μ_2 < 0 (same as H_a: μ_1 < μ_2) * H_a: μ_1 - μ_2 0 (same as H_a: μ_1 > μ_2) * H_a: μ_1 - μ_2 ≠ 0 (same as H_a: μ_1 ≠ μ_2) For a paired t-test, the hypotheses are H_0: μ_d = 0, and one of: * H_a: μ_d < 0, * H_a: μ_d > 0, * H_0: μ_0 ≠ 0. For ANOVA, H_0: μ_0 = μ_2 = ... = μ_k, and H_a:not all μ's are equal" height="311" loading="lazy" src="http://phhp-faculty-cantrell.sites.m...c-q_table2.png" title="In a Two-Sample t-test, the hypotheses are: H_0: μ_1 - μ_2 = 0 (or H_0: μ_1 = μ_2), and one of: * H_a: μ_1 - μ_2 < 0 (same as H_a: μ_1 < μ_2) * H_a: μ_1 - μ_2 > 0 (same as H_a: μ_1 > μ_2) * H_a: μ_1 - μ_2 ≠ 0 (same as H_a: μ_1 ≠ μ_2) For a paired t-test, the hypotheses are H_0: μ_d = 0, and one of: * H_a: μ_d < 0, * H_a: μ_d > 0, * H_0: μ_0 ≠ 0. For ANOVA, H_0: μ_0 = μ_2 = ... = μ_k, and H_a:not all μ's are equal" width="610">

Step 2: Check Conditions and Summarize the Data Using a Test Statistic

We need to check that the conditions under which the test can be reliably used are met.

For the Paired t-test (as a special case of a one-sample t-test), the conditions are:

- The sample of differences is random (or at least can be considered so in context).
- We are in one of the three situations marked with a green check mark in the following table:



For the Two-Sample t-test, the conditions are:

- Two samples are independent and random
- One of the following two scenarios holds:
  - Both populations are normal
  - Populations are not normal, but large sample size (>30)

For an ANOVA, the conditions are:

- The samples drawn from each of the populations being compared are independent.
- The response variable varies normally within each of the populations being compared. As is often the case, we do not have to worry about this assumption for large sample sizes.
- The populations all have the same standard deviation.

Now we summarize the data using a test statistic.

- Although we will not be calculating these test statistics by hand, we will review the formulas for each test statistic here.

For the Paired t-test the test statistic is:

$$t = \frac{\bar{y}_d - 0}{s_d/\sqrt{n}}$$

For the Two-Sample t-test assuming equal variances the test statistic is:

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

For the Two-Sample t-test assuming unequal variances the test statistic is:

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For an ANOVA the test statistic is:

$$F = \frac{\text{VARIATION AMONG SAMPLE MEANS}}{\text{VARIATION WITHIN GROUPS}}$$

Step 3: Finding the p-value of the test

Use statistical software to determine the p-value.

- The p-value is the probability of getting data like those observed (or even more extreme) assuming that the null hypothesis is true, and is calculated using the null distribution of the test statistic.
- The p-value is a measure of the evidence against H_0.
- The smaller the p-value, the more evidence the data present against H_0.

The p-values for three C→Q tests are obtained from the output.

Step 4: Making conclusions

Conclusions about the significance of the results:

- If the p-value is small, the data present enough evidence to reject H_o (and accept H_a).
- If the p-value is not small, the data do not provide enough evidence to reject H_0.
- To help guide our decision, we use the significance level as a cutoff for what is considered a small p-value. The significance cutoff is usually set at .05, but should not be considered inviolable.

Conclusions should always be stated in the context of the problem and can all be written in the basic form below:

- There (IS or IS NOT) enough evidence that there is an association between (X) and (Y). Where X and Y should be given in context.

Following the test…

- For a paired t-test, a **95% confidence interval** for μ_d can be very insightful after a test has rejected the null hypothesis, and can also be used for testing in the two-sided case.
- For a two-sample t-test, a **95% confidence interval** for $\mu_1 - \mu_2$ can be very insightful after a test has rejected the null hypothesis, and can also be used for testing in the two-sided case.
- If the ANOVA F-test has rejected the null hypothesis, looking at the **confidence intervals** for the population means that are in the output can provide visual insight into why the H_0 was rejected (i.e., which of the means differ).

Non-parametric Alternatives

- For a Paired t-test we might investigate using the Wilcoxon Signed-Rank test or the Sign test.
- For a Two-Sample t-test we might investigate using the Wilcoxon Rank-Sum test (Mann-Whitney U test).
- For an ANOVA we might investigate using the Kruskal-Wallis test.

Case Q→Q

CO-4: Distinguish among different measurement scales, choose the appropriate descriptive and inferential statistical methods based on these distinctions, and interpret the results.

Learning Objectives

LO 4.35: For a data analysis situation involving two variables, choose the appropriate inferential method for examining the relationship between the variables and justify the choice.

Learning Objectives

LO 4.36: For a data analysis situation involving two variables, carry out the appropriate inferential method for examining relationships between the variables and draw the correct conclusions in context.

CO-5: Determine preferred methodological alternatives to commonly used statistical methods when assumptions are not met.

Review: From UNIT 1

- Case Q-Q
  - Scatterplots
  - Linear Relationships – Correlation
  - Linear Relationships – Linear Regression

Video

Video: Case Q→Q (60:27)

Related SAS Tutorials

- 9A – (3:53) Basic Scatterplots
- 9B – (2:29) Grouped Scatterplots
- 9C – (3:46) Pearson's Correlation Coefficient
- 9D – (3:00) Simple Linear Regression – EDA
- 9E – (5:59) Simple Linear Regression (Inference)

Related SPSS Tutorials

- 9A – (2:38) Basic Scatterplots
- 9B – (2:54) Grouped Scatterplots
- 9C – (3:35) Pearson's Correlation Coefficient
- 9D – (2:53) Simple Linear Regression – EDA
- 9E – (7:07) Simple Linear Regression (Inference)

Introduction

In inference for relationships, so far we have learned inference procedures for both cases C→Q and C→C from the role/type classification table below.

The last case to be considered in this course is case Q→Q, where both the explanatory and response variables are quantitative. (Case Q→C requires statistical methods that go beyond the scope of this course, one of which is logistic regression).

		Response	
		Categorical	Quantitative
Explanatory	Categorical	✓C →C	✓C →Q
	Quantitative	Q →C	**Q →Q**

For case Q→Q, we will learn the following tests:

	Dependent Samples	Independent Samples
Standard Test(s)	• Not Covered (Longitudinal Data Analysis, etc.)	• Test for Significance of Pearson's Correlation Coefficient • Test for Significance of the Slope in Linear Regression
Non-Parametric Test(s)		• Test for Significance of Spearman's Rank Correlation

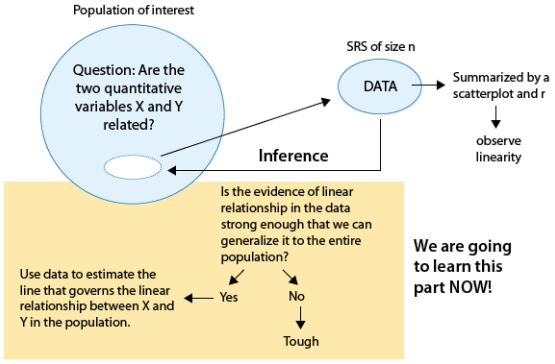
In the Exploratory Data Analysis section, we examined the relationship between sample values for two quantitative variables by looking at a scatterplot and **if the relationship was linear**, we supplemented the scatterplot with the correlation coefficient r and the linear regression equation. We discussed the regression equation but made no attempt to claim that the relationship observed in the sample necessarily held for the larger population from which the sample originated.

Now that we have a better understanding of the process of statistical inference, we will discuss a few methods for inferring something about the relationship between two quantitative variables in an entire population, based on the relationship seen in the sample.

> In particular, we will focus on **linear** relationships and will answer the following questions:
>
> - **Is the correlation coefficient different from zero in the population, or could it be that we obtained the result in the data just by chance?**
> - **Is the slope different from zero in the population, or could it be that we obtained the result in the data just by chance?**

If we satisfy the assumptions and conditions to use the methods, we can estimate the slope and correlation coefficient for our population and conduct hypothesis tests about these parameters.

For the standard tests, the tests for the slope and the correlation coefficient are equivalent; they will always produce the same p-value and conclusion. This is because they are directly related to each other.



In this section, we can state our **null and alternative hypotheses as:**

Ho: There is no relationship between the two quantitative variables X and Y.

Ha: There is a relationship between the two quantitative variables X and Y.

Pearson's Correlation Coefficient

Learning Objectives

LO 4.45: In a given context, set up the appropriate null and alternative hypotheses for examining the relationship between two quantitative variables.

Learning Objectives

LO 4.46: In a given context, determine the appropriate standard method for examining the relationship between two quantitative variables interpret the results provided in the appropriate software output in context.

What we know from Unit 1:

- r only measures the LINEAR association between two quantitative variables X and Y
- $-1 \leq r \leq 1$
- If the relationship is linear then:

  r = 0 implies no relationship between X and Y (note this is our null hypothesis!!)

  r > 0 implies a positive relationship between X and Y (as X increases, Y also increases)

  r < 0 implies a negative relationship between X and Y (as X increases, Y decreases)

Now here are the steps for hypothesis testing for Pearson's Correlation Coefficient:

Step 1: State the hypotheses If we consider the above information and our null hypothesis,

Ho: There is no relationship between the two quantitative variables X and Y,

Before we can write this using correlation, we must define the population correlation coefficient. In statistics, we use the greek letter ρ (rho) to denote the population correlation coefficient. Thus if there is no relationship between the two quantitative variables X and Y in our population, we can see that this hypothesis is equivalent to

Ho: $\rho = 0$ (rho = 0).

The alternative hypothesis will be

Ha: $\rho \neq 0$ (rho is not equal to zero).

however, one sided tests are possible.

Step 2: Obtain data, check conditions, and summarize data

(i) The sample should be random with independent observations (all observations are independent of all other observations).

(ii) The relationship should be reasonably linear which we can check using a scatterplot. Any clearly non-linear relationship should not be analyzed using this method.

(iii) To conduct this test, both variables should be normally distributed which we can check using histograms and QQ-plots. Outliers can cause problems.

Although there is an intermediate test statistic, in effect, the value of r itself serves as our test statistic.

Step 3: Find the p-value of the test by using the test statistic as follows

We will rely on software to obtain the p-value for this test. We have seen this p-value already when we calculated correlation in Unit 1.

Step 4: Conclusion

As usual, we use the magnitude of the p-value to draw our conclusions. A small p-value indicates that the evidence provided by the data is strong enough to reject Ho and conclude (beyond a reasonable doubt) that the two variables are related ($\rho \neq 0$).

In particular, if a significance level of 0.05 is used, we will reject Ho if the p-value is less than 0.05.

Confidence intervals can be obtained to estimate the true population correlation coefficient, ρ (rho), however, we will not compute these intervals in this course. You could be asked to interpret or use a confidence interval which has been provided to you.

Non-Parametric Alternative: Spearman's Rank Correlation

Learning Objectives

LO 5.1: For a data analysis situation involving two variables, determine the appropriate alternative (non-parametric) method when assumptions of our standard methods are not met.

Learning Objectives

LO 5.2: Recognize situations in which Spearman's rank correlation is a more appropriate measure of the relationship between two quantitative variables

We will look at one non-parametric test in case Q→Q. Spearman's rank correlation uses the same calculations as for Pearson's correlation coefficient except that it uses the ranks instead of the original data. This test is useful when there are outliers or when the variables do not appear to be normally distributed.

- This measure and test are most useful **when the relationship between X and Y is nonlinear and either non-increasing or non-decreasing.**
- **If the relationship has both increasing and decreasing components, Spearman's rank correlation is not usually helpful** as a measure of correlation.

This measure behaves similarly to r in that:

- it ranges from -1 to 1
- a value of 0 implies no relationship
- positive values imply a positive relationship
- negative values imply a negative relationship.

Now an example:

EXAMPLE: IQ vs. Cry Count

A method for predicting IQ as soon as possible after birth could be important for early intervention in cases such as brain abnormalities or learning disabilities. It has been thought that greater infant vocalization (for instance, more crying) is associated with higher IQ. In 1964, a study was undertaken to see if IQ at 3 years of age is associated with amount of crying at newborn age. In the study, 38 newborns were made to cry after being tapped on the foot and the number of distinct cry vocalizations within 20 seconds was counted. The subjects were followed up at 3 years of age and their IQs were measured.

Data: SPSS format, SAS format, Excel format

Response Variable:

- IQ at three years of age

Explanatory Variable:

- Newborn cry count in 20 seconds

Results:

Step 1: State the hypotheses

The hypotheses are:

Ho: There is no relationship between newborn cry count and IQ at three years of age

Ha: There is a relationship between newborn cry count and IQ at three years of age

Steps 2 & 3: Obtain data, check conditions, summarize data, and f**ind the p-value**

(i) To the best of our knowledge the subjects are independent.

(ii) The scatterplot shows a relationship that is reasonably linear although not very strong.

(iii) The histograms and QQ-plots for both variables are slightly skewed right. We would prefer more symmetric distributions; however, the skewness is not extreme so we will proceed with caution.

Pearson's correlation coefficient is 0.402 with a p-value of 0.012.

Spearman's rank correlation is 0.354 with a p-value of 0.029.

Step 4: Conclusion

Based upon the scatterplot and correlation results, there is a statistically significant, but somewhat weak, positive correlation between newborn cry count and IQ at age 3.

SPSS Output for tests

SAS Output, SAS Code

Simple Linear Regression

Learning Objectives

LO 4.46: In a given context, determine the appropriate standard method for examining the relationship between two quantitative variables interpret the results provided in the appropriate software output in context.

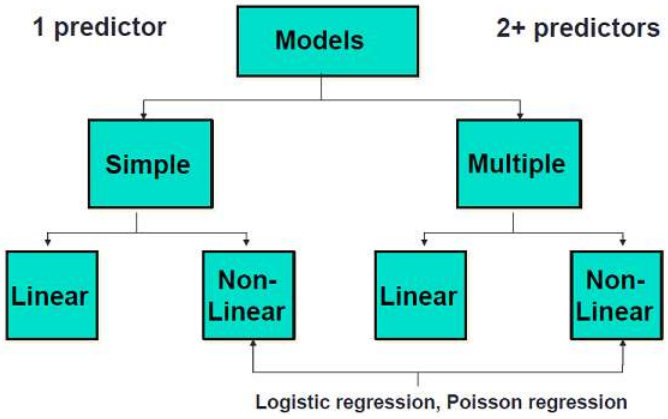
In Unit 1, we discussed the least squares method for estimating the regression line and used software to obtain the slope and intercept of the linear regression equation. These estimates can be considered as the sample statistics which estimate the true population slope and intercept.

Now we will formalize simple linear regression which will require some additional notation.

A **regression model expresses two essential ingredients**:

- a tendency of the response variable Y to vary with the explanatory variable X in a **systematic** fashion (**deterministic**)
- a stochastic **scattering of points around the curve** of statistical relationship (**random**)

Regression is a vast subject which handles a wide variety of possible relationships.



All regression methods begin with a theoretical model which specifies the form of the relationship and includes any needed assumptions or conditions. Now we will introduce a more "statistical" definition of the regression model and define the parameters in the population.

Simple Linear Regression Model:

We will use a different notation here than in the beginning of the semester. Now we use regression model style notation.

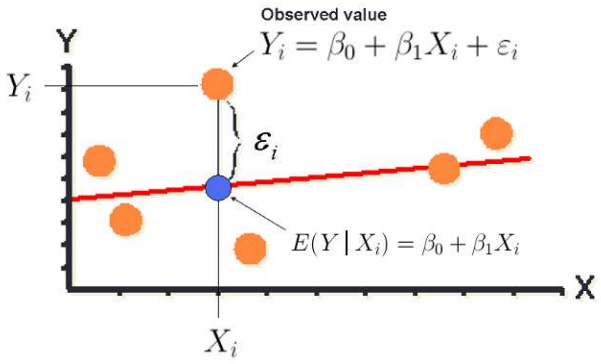
We **assume the relationship in the population is linear** and therefore our regression model can be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

- The parameter β_0 (beta_zero) is the intercept (in the population) and is the average value of Y when X = 0
- The parameter β_1 (beta_1) is the slope (in the population) and is the change in the average Y for each 1 unit increase in X.
- X_i is the value of the explanatory variable for the i-th subject
- Y_i is the value of the response variable for the i-th subject
- ε_i (epsilon_i) is the error term for the i-th subject
- the error terms are assumed to be
  - **normally distributed with mean zero** (check with histogram and QQ-plot of the residuals)
  - **constant variance** (check with scatterplot of Y vs. X for simple linear regression)
  - statistically independent (difficult to check, be sure to have independent observations in the data, different methods are required for dependent observations!)

The following picture illustrates the components of this model.



Each orange dot represents an individual observation in the scatterplot. Each observed value is modeled using the previous equation.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The red line is the true linear regression line. The blue dot represents the predicted value for a particular X value and illustrates that our predicted value only estimates the mean, average, or expected value of Y at that X value.

The error for an individual is expected and is due to the variation in our data. In the previous illustration, it is labeled with ε_i (epsilon_i) and denoted by a bracket which gives the distance between the orange dot for the observed value and the blue dot for the predicted value for a particular value of X. In practice, we cannot observe the true error for an individual but we will be able to estimate them using the residuals, which we will soon define mathematically.

The **regression line represents the average Y for a given X** and can be expressed as in symbols as the **expected value of Y for a given X, E(Y|X) or Y-hat.**

$$E(Y|X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

In Unit 1, we used **a** to represent the intercept and **b** to represent the slope that we estimated from our data.

In formal regression procedures, we commonly use beta to represent the population parameter and beta-hat to represent the **parameter estimate**.

These **parameter estimates**, which are sample statistics estimated from our data, are also sometimes referred to as the **coefficients** using algebra terminology.

For each observation in our dataset, we also have a **residual** which is defined as the difference between the observed value and the predicted value for that observation.

$$\text{residual}_i = Y_i - \hat{Y}_i$$

The residuals are used to check our assumptions of normality and constant variance.

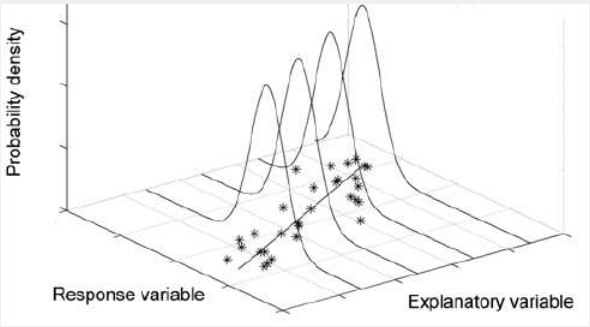
In effect, since we have a quantitative response variable, we are still comparing population means. However, now we must do so for EVERY possible value of X. We want to know if the distribution of Y is the same or different over our range of X values.

This idea is illustrated (including our assumption of normality) in the following picture which shows a case where the distribution of Y is changing as the values of the explanatory variable X change. This change is reflected by only a shift in means since we assume normality and constant variation of Y for all X.

The method used is mathematically equivalent to ANOVA but our interpretations are different due to the quantitative nature of our explanatory variable.

This image shows a scatterplot and regression line on the X-Y plane – as if flat on a table. Then standing up – in the vertical axis – we draw normal curves centered at the regression line for four different X-values – with X increasing for each.

The center of the distributions of the normal distributions which are displayed shows an increase in the mean but constant variation.



The idea is that the model assumes a normal distribution is a good approximation for how the Y-values will vary around the regression line for a particular value of X.

Coefficient of Determination

Learning Objectives

LO 4.47: For simple linear regression models, interpret the coefficient of determination in context.

There is one additional measure which is often of interest in linear regression, **the coefficient of determination, R^2** which, for simple linear regression is simply the square of the correlation coefficient, r.

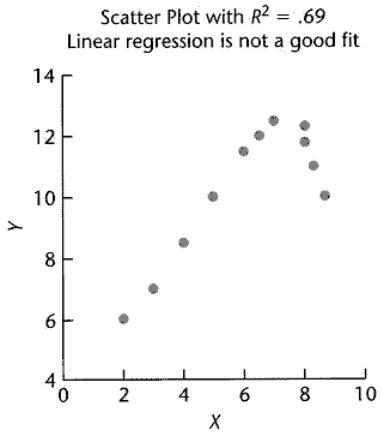
The value of R^2 is interpreted as **the proportion of variation in our response variable Y, which can be explained by the linear regression model using our explanatory variable X.**

Important Properties of R^2

- $0 \leq R^2 \leq 1$
- $R^2 = 0$ implies the model explains none of the variation in Y.
- $R^2 = 1$ implies the model explains all of the variation in Y (perfect fit, very unlikely with data)

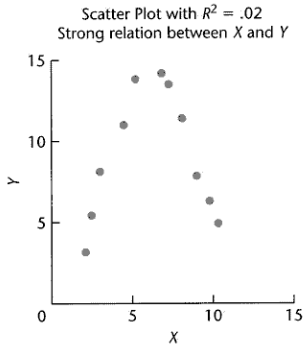
A **large R^2 may or MAY NOT mean that the model fits our data well.**

The image below illustrates data with a fairly large R^2 yet the model does not fit the data well.



A **small R^2 may or MAY NOT mean that there is no relationship** between X and Y – we must be careful as the relationship that exists may simply not be specified in our model – currently a simple linear model.

The image below illustrates data with a very small R^2 yet the true relationship is very strong.



Test Procedure for the Slope in Simple Linear Regression

Now we move into our formal test procedure for simple linear regression.
A small R2 may or MAY NOT mean that there is no relationship between X and Y – we must be careful as the relationship that exists may simply not be specified in our model – currently a simple linear model. The image below illustrates data with a very small R2 yet the true relationship is very strong.

Step 1: State the hypotheses

In order to test the hypothesis that

Ho: There is no relationship between the two quantitative variables X and Y,

assuming our model is correct (a linear model is sufficient), we can write the above hypothesis as

Ho: $\beta_1 = 0$ (Beta_1 = 0, the slope of our linear equation = 0 in the population).

The alternative hypothesis will be

Ha: $\beta_1 \neq 0$ (Beta_1 is not equal to zero).

Step 2: Obtain data, check conditions, and summarize data

(i) The sample should be random with independent observations (all observations are independent of all other observations).

(ii) The relationship should be linear which we can check using a scatterplot.

(iii) The residuals should be reasonably normally distributed with constant variance which we can check using the methods discussed below.

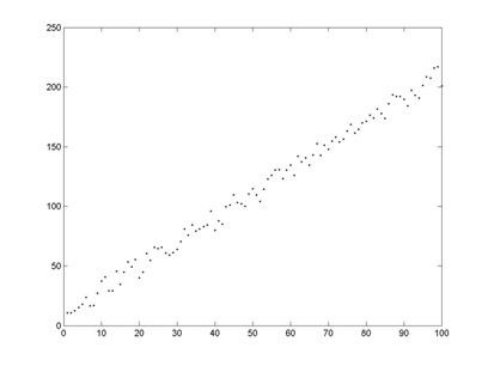
Normality: Histogram and QQ-plot of the residuals.

Constant Variance: Scatterplot of Y vs. X and/or a scatterplot of the residuals vs. the predicted values (Y-hat). We would like to see random scatter with no pattern and approximately the same spread for all values of X.

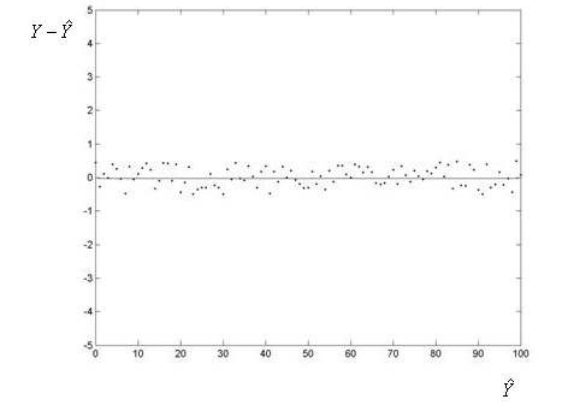
Large outliers which fall outside the pattern of the data can cause problems and exert undue influence on our estimates. We saw in Unit 1 that one observation which is far away on the x-axis can have an large impact on the values of the correlation and slope.

Here are two examples each using the two plots mentioned above.

Example 1: Has constant variance (homoscedasticity)

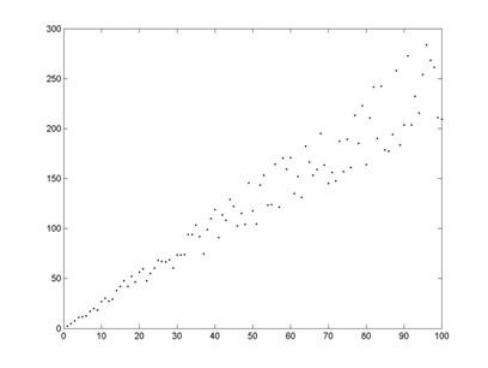


Scatterplot of Y vs. X (above)

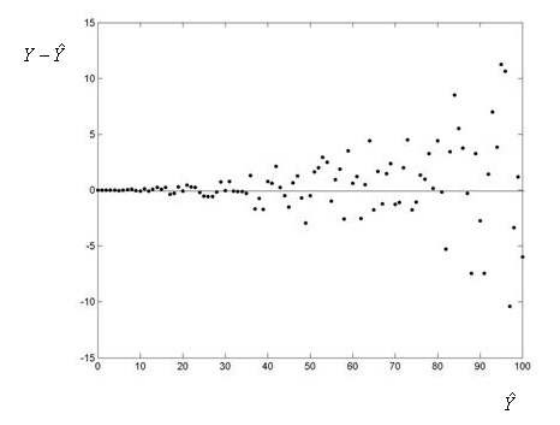


Scatterplot of residuals vs. predicted values (above)

Example 2: Does not have constant variance (heteroscedasticity)



Scatterplot of Y vs. X (above)



Scatterplot of residuals vs. predicted values (above)

The **test statistic** is similar to those we have studied for other t-tests:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}_{\hat{\beta}_1}}$$

where

$\text{SE}_{\hat{\beta}_1}$ = standard error of $\hat{\beta}_1$.

Both of these values, along with the test statistic, are provided in the output from the software.

Step 3: Find the p-value of the test by using the test statistic as follows

Under the null hypothesis, the test statistic follows a t-distribution with n-2 degrees of freedom. We will rely on software to obtain the p-value for this test.

Step 4: Conclusion

As usual, we use the magnitude of the p-value to draw our conclusions. A small p-value indicates that the evidence provided by the data is strong enough to reject Ho and we would conclude there is enough evidence that hat slope in the population is not zero and therefore the two variables are related. In particular, if a significance level of 0.05 is used, we will reject Ho if the p-value is less than 0.05.

Confidence intervals will also be obtained in the software to estimate the true population slope, β_1 (beta_1).

EXAMPLE: IQ vs. Cry Count

A method for predicting IQ as soon as possible after birth could be important for early intervention in cases such as brain abnormalities or learning disabilities. It has been thought that greater infant vocalization (for instance, more crying) is associated with higher IQ. In 1964, a study was undertaken to see if IQ at 3 years of age is associated with amount of crying at

newborn age. In the study, 38 newborns were made to cry after being tapped on the foot and the number of distinct cry vocalizations within 20 seconds was counted. The subjects were followed up at 3 years of age and their IQs were measured.

Data: SPSS format, SAS format, Excel format

Response Variable:

- IQ at three years of age

Explanatory Variable:

- Newborn cry count in 20 seconds

Results:

Step 1: State the hypotheses

The hypotheses are:

Ho: There is no (linear) relationship between newborn cry count and IQ at three years of age

Ha: There is a (linear) relationship between newborn cry count and IQ at three years of age

Steps 2 & 3: Obtain data, check conditions, summarize data, and f**ind the p-value**

(i) To the best of our knowledge the subjects are independent.

(ii) The scatterplot shows a relationship that is reasonably linear although not very strong.

(iii) The histogram and QQ-plot of the residuals are both reasonably normally distributed. The scatterplots of Y vs. X and the residuals vs. the predicted values both show no evidence of non-constant variance.

The estimated regression equation is

$$\hat{IQ} = 90.76 + 1.54(\text{cry count})$$

The parameter estimate of the slope is 1.54 which means that for each 1-unit increase in cry count, the average IQ is expected to increase by 1.54 points.

The standard error of the estimate of the slope is 0.584 which give a test statistic of 2.63 in the output and using unrounded values from the output and the formula:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}_{\hat{\beta}_1}} = \frac{1.536 - 0}{0.584} = 2.63\ .$$

The p-value is found to be 0.0124. Notice this exactly the same as we obtained for this data for our test of Pearson's correlation coefficient. These two methods are equivalent and will always produce the same conclusion about the statistical significance of the linear relationship between X and Y.

The 95% confidence interval for β_1 (beta_1) given in the output is (0.353, 2.720).

This regression model has coefficient of determination of R^2 = 0.161 which means that 16.1% of the variation in IQ score at age three can be explained by our linear regression model using newborn cry count. This confirms a relatively weak relationship as we found in our previous example using correlations (Pearson's correlation coefficient and Spearmans' rank correlation).

Step 4: Conclusion

Conclusion of the test for the slope: Based upon the scatterplot and linear regression analysis, since the relationship is linear and the p-value = 0.0124, there is a statistically significant positive linear relationship between newborn cry count and IQ at age 3.

Interpretation of R-squared: Based upon our $R^2$and scatterplot, the relationship is somewhat weak with only 16.1% of the variation in IQ score at age three being explained by our linear regression model using newborn cry count.

Interpretation of the slope: For each 1-unit increase in cry count, the population mean IQ is expected to increase by 1.54 points, however, the 95% confidence interval suggests this value could be as low as 0.35 points to as high as 2.72 points.

SPSS Output for tests

SAS Output, SAS Code

EXAMPLE: Gestation vs. Longevity in Animals

We return to the data from an earlier activity (Learn By Doing – Correlation and Outliers (Software)). The average gestation period, or time of pregnancy, of an animal is closely related to its longevity, the length of its lifespan. Data on the average gestation period and longevity (in captivity) of 40 different species of animals have been recorded. Here is a summary of the variables in our dataset:

- **animal:** the name of the animal species.
- **gestation:** the average gestation period of the species, in days.
- **longevity:** the average longevity of the species, in years.

In this case, whether we include the outlier or not, there is a problem of non-constant variance. You can clearly see that, in general, as longevity increases, the variation of gestation increases.

This data is not a particularly good candidate for simple linear regression analysis (without further modification such as transformations or the use of alternative methods).

Pearson's correlation coefficient (or Spearman's rank correlation), may still provide a reasonable measure of the strength of the relationship, which is clearly a positive relationship from the scatterplot and our previous measure of correlation.

Output – Contains scatterplots with linear equations and LOESS curves (running average) for the dataset with and without the outlier. Pay particular attention to the problem with non-constant variance seen in these scatterplots.

EXAMPLE: Insurance Premiums

The data used in the analysis provided below contains the monthly premiums, driving experience, and gender for a random sample of drivers.

To analyze this data, we have looked at males and females as two separate groups and estimated the correlation and linear regression equation for each gender. We wish to predict the monthly premium using years of driving experience.

Use this output for additional practice with these concepts. For each gender consider the following:

- Are the assumptions satisfied?
- Is the correlation statistically significant? Is it positive or negative? Weak or strong?
- Is the slope statistically significant? What does the slope mean in context? What is the confidence interval for the slope?
- What is R^2 and what does it mean in context?

SPSS Output

SAS Output, SAS Code

Wrap-Up (Inference for Relationships)

Video

Video: Full Course Overview & Summary (68:32)

Learn By Doing: Supplemental Examples and Exercises for Unit 4B
(Non-interactive Version)

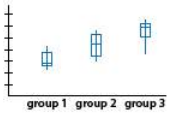
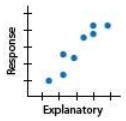
We've just completed the part of the course about the inferential methods for relationships between variables. The overall goal of inference for relationships is to assess whether the observed data provide evidence of a significant relationship between the two variables (i.e., a true relationship that exists in the population).

Much like the unit about relationships in the Exploratory Data Analysis (EDA) unit, this part of the course was organized according to the role and type classification of the two variables involved.

However, unlike the EDA unit , when it comes to inferential methods, we further distinguished between three sub-cases in case C→Q, so essentially we covered 5 cases in total.

The following very detailed role-type classification table summarizes both EDA and inference for the relationship between variables:

Role-Type Classification Table

Explanatory	Response: Categorical	Response: Quantitative
Categorical	**C → C** **To Visualize** 2-Way Table (Outcome A / Outcome B / Outcome C; Group 1 / Group 2 / Group 3) **Numerical Summary** Conditional Percentages **Formal Inference** Chi-Square test for Independance	**C → Q** **To Visualize** Side-by-side Boxplots  **Numerical Summary** Descriptive Statistics **Formal Inference** **2 independant** samples: Two-Sample t-test **2 dependant** samples: Paired t-test **> 2 independant** samples: ANOVA **> 2 dependant** samples: Not covered in the course
Quantitative	**Q → C** **Logistic Regression** Not covered in this course	**Q → Q** **To Visualize** Scatterplot  **Numerical Summaries** Correlation Coefficient **Formal Inference** Regression line. Significance test for the linear relationship (t-test for the slope).

Case C-Q

Here is a summary of the tests for the scenario where k = 2.

Independent Samples (More Emphasis)	Dependent Samples (Less Emphasis)
Standard Tests • **Two Sample T-Test Assuming Equal Variances** • **Two Sample T-Test Assuming Unequal Variances** Non-Parametric Test • Mann-Whitney U (or Wilcoxon Rank-Sum) Test	**Standard Test** • **Paired T-Test** Non-Parametric Tests • Sign Test • Wilcoxon Signed-Rank Test

Here is a summary of the tests for the scenario where k > 2.

Independent Samples (Only Emphasis)	Dependent Samples (Not Discussed)
Standard Tests • **One-way ANOVA (Analysis of Variance)** Non-Parametric Test • Kruskal–Wallis One-way ANOVA	*Standard Test* • *Repeated Measures ANOVA (or similar)*

Case C-C

Independent Samples (Only Emphasis)	Dependent Samples (Not Discussed)
Standard Tests • **Continuity Corrected Chi-square Test for Independence (2×2 case)** • **Chi-square Test for Independence (RxC case)** Non-Parametric Test • Fisher's exact test	*Standard Test* • *McNemar's Test – 2×2 Case*

Case Q-Q

Independent Samples (Only Emphasis)	Dependent Samples (Not Discussed)
Standard Tests • **Test for Significance of Pearson's Correlation Coefficient** • **Test for Significance of the Slope in Linear Regression** Non-Parametric Test • Test for Significance of Spearman's Rank Correlation	*Standard Test* • *Not Covered (Longitudinal Data Analysis, etc.)*

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Biostatistics - Open Learning Textbook

Webpages: 55

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 81.8% (45 pages)
- Undeclared: 18.2% (10 pages)

By Page

- Biostatistics - Open Learning Textbook - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - Preliminaries - *CC BY-NC-SA 4.0*
    - Role of Biostatistics - *CC BY-NC-SA 4.0*
    - The Big Picture - *CC BY-NC-SA 4.0*
    - Types of Variables - *CC BY-NC-SA 4.0*
    - What is Data? - *Undeclared*
  - Unit 1: Exploratory Data Analysis - *CC BY-NC-SA 4.0*
    - Case C-C - *CC BY-NC-SA 4.0*
    - Case C-Q - *CC BY-NC-SA 4.0*
    - Case Q-Q - *CC BY-NC-SA 4.0*
    - Causation - *CC BY-NC-SA 4.0*
    - One Categorical Variable - *CC BY-NC-SA 4.0*
    - One Quantitative Variable: Introduction - *CC BY-NC-SA 4.0*
    - Role-Type Classification - *CC BY-NC-SA 4.0*
    - Summary (Unit 1) - *CC BY-NC-SA 4.0*
  - Unit 2: Producing Data - *CC BY-NC-SA 4.0*
    - Causation and Experiments - *CC BY-NC-SA 4.0*
    - Causation and Observational Studies - *CC BY-NC-SA 4.0*
    - Designing Studies - *CC BY-NC-SA 4.0*
    - Sample Surveys - *CC BY-NC-SA 4.0*
    - Sampling - *CC BY-NC-SA 4.0*
    - Summary (Unit 2) - *CC BY-NC-SA 4.0*
  - Unit 3A: Probability - *CC BY-NC-SA 4.0*
    - Basic Probability Rules - *CC BY-NC-SA 4.0*
    - Conditional Probability and Independence - *CC BY-NC-SA 4.0*
    - Introduction to Probability - *CC BY-NC-SA 4.0*
    - Summary (Unit 3) - *CC BY-NC-SA 4.0*
  - Unit 3B: Random Variables - *CC BY-NC-SA 4.0*
    - Binomial Random Variables - *CC BY-NC-SA 4.0*
    - Continuous Random Variables - *CC BY-NC-SA 4.0*
    - Discrete Random Variables - *CC BY-NC-SA 4.0*
    - Normal Random Variables - *CC BY-NC-SA 4.0*
    - Summary (Unit 3B - Random Variables) - *CC BY-NC-SA 4.0*
  - Unit 3B: Sampling Distribution - *CC BY-NC-SA 4.0*
    - Sampling Distribution of the Sample Mean, x-bar - *CC BY-NC-SA 4.0*
    - Sampling Distribution of the Sample Proportion, p-hat - *CC BY-NC-SA 4.0*
    - Summary (Unit 3B - Sampling Distributions) - *CC BY-NC-SA 4.0*
  - Unit 4A: Introduction to Statistical Inference - *CC BY-NC-SA 4.0*
    - Estimation - *CC BY-NC-SA 4.0*
    - Hypothesis Testing - *CC BY-NC-SA 4.0*
    - Wrap-Up (Inference for One Variable) - *CC BY-NC-SA 4.0*
  - Unit 4B: Inference for Relationships - *CC BY-NC-SA 4.0*
    - Case C→C - *CC BY-NC-SA 4.0*
    - Case C→Q - *CC BY-NC-SA 4.0*
    - Case Q→Q - *CC BY-NC-SA 4.0*
    - Wrap-Up (Inference for Relationships) - *CC BY-NC-SA 4.0*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*