



# INVESTIGATING STATISTICAL CONCEPTS, APPLICATIONS, AND METHODS

Beth L. Chance

Allan J. Rossman

*Third Edition*



### Acknowledgements

We would like to thank the following reviewers for their continued feedback and suggestions. They have greatly improved the quality of these materials. We would like to especially thank Julie Clark of Hollins University for her gracious and extensive attention to detail.

Doug Andrews, Wittenberg University; Julie Clark, Hollins University; David Cruz-Uribe, Trinity College; Jo Hardin, Pomona College; Allen Hibbard, Central College; Sharon Lane-Getaz, St. Olaf University; Tom Linton, Central College; Mark Mills, Central College; Paul Roback, St. Olaf College; Soma Roy, Cal Poly – San Luis Obispo; Michael Stob, Calvin College

Cover Design: Danna Currie, Grover Beach

Initial Solutions to Investigations: Chelsea Lofland, based on classroom notes

Initial Exercise Compilation and Solutions: Julie Clark

Copy-editing Assistance: Virginia Burroughs

Companion Website: [www.rossmchance.com/iscam3/](http://www.rossmchance.com/iscam3/)

Online Resources include solutions to investigations, exercises, and additional investigations.

Also see Resources for Instructors.

Copyright © August 2015 Beth Chance and Allan Rossman  
San Luis Obispo, California

ALL RIGHTS RESERVED. No part of this work covered by the copyright hereon may be produced or - used in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping, Web distribution, information storage retrieval systems, or in any other manner – without the written permission of the authors.

**Table of Contents**

To The Student.....	3
Investigation A: Traffic Fatalities And Federal Speed Limits.....	4
Investigation B: Random Babies .....	12
Chapter 1: Analyzing One Categorical Variable .....	19
Section 1: Analyzing a Process Probability .....	20
Section 2: Normal Approximations for Sample Proportions.....	62
Section 3: Sampling From a Finite Population .....	91
Chapter 3: Comparing Two Proportions.....	180
Section 1: Comparing Two Population Proportions .....	181
Section 2: Types of Studies.....	196
Section 3: Comparing Two Treatment Probabilities .....	205
Section 4: Other Statistics .....	221
Chapter 4: Comparisons With Quantitative Variables.....	248
Section 1: Comparing Groups – Quantitative Reponse .....	249
Section 2: Comparing Two Population Means .....	252
Section 3: Comparing Two Treatment Means .....	264
Section 4: Matched Pairs Designs.....	281
Chapter 5: Comparing Several Populations, Exploring Relationships .....	312
Section 1: Two Categorical Variables .....	313
Section 2: Comparing Several Population Means .....	330
Section 3: Relationships Between Quantitative Variables.....	346
Section 4: Inference For Regression .....	374
Index 417	

## TO THE STUDENT

Statistics is a mathematical science.

Although this is a very short sentence, perhaps a self-evident one, and certainly one of the shortest that you will find in this book, we want to draw your attention to several things about it:

- We use the singular “is” and not the plural “are.” It is certainly grammatically correct and more common usage to say “statistics are...”, but that use of the term refers to statistics as numerical values. In this sentence we mean statistics as a field of study, one that has its own concepts and techniques, and one that can be exciting to study and practice.
- We use “mathematical” as an adjective. Statistics certainly makes use of much mathematics, but it is a separate discipline and not a branch of mathematics. Many, perhaps most, of the concepts and methods in statistics are mathematical in nature, but there are also many that do not involve mathematics. You will see an example of this early in the book as you study the difference between observational studies and controlled experiments. You will find that even in cases where the mathematical aspects of two situations may be identical, the scope of one’s conclusions depends crucially on how the data were collected, a statistical rather than a mathematical consideration.
- We use the noun “science.” Statistics is the science of gaining insight from data. Data are (notice the plural here) pieces of information (often but not always numerical) gathered on people or objects or processes. The science of statistics involves all aspects of inquiry about data. Well-designed studies begin with a research question or hypothesis, devise a plan for collecting data to address that issue, proceed to gather the data and analyze them, and then often make inferences about how the findings generalize beyond the particular group being studied. Statistics concerns itself with all phases of this process and therefore encompasses the scientific method.

In these materials, our goal is to introduce you to this practice of statistics, to help you think about the applications of statistics and to study the mathematical underpinnings of the statistical methods. Most of all, we hope you will find fun and engaging examples. Statistics is a vitally important subject, and also fun to study and practice, largely because it brings you into contact with all kinds of interesting questions. You will analyze data from medical studies, legal cases, psychology experiments, sociological studies, and many other contexts. To paraphrase the late statistician John Tukey, “the best thing about statistics is that it allows you to play in everyone else’s backyard.” You never know what you might learn in a statistics class!

One of the first features you will notice about these materials is that you will play the active role of investigator. You will read about an actual study and consider the research question, and then we will lead you to discover and apply the appropriate tools for carrying out the analysis. A primary reason for the investigative nature of these materials is that we strongly believe that you will better understand and retain the concepts if you build your own knowledge and are engaged in the context. Be sure to also pay attention to the Study Conclusions to see how to effectively convey statistical information and the Practice Problems for testing your understanding. Though you will only scratch the surface of the statistical methods used in practice, you will learn fundamental concepts (such as variability, randomness, confidence, and significance) that are an integral part of many statistical analyses. A distinct emphasis will be the focus on how the data are collected and how this determines the scope of conclusions that you can draw from the data.

## Investigation A: Traffic Fatalities and Federal Speed Limits

These first two investigations give you a very brief introduction to some big ideas for the course. Some of you will have seen some of these ideas before and can use the investigations to refresh your memory. Some of the ideas may be new and you will see them again in later chapters. For now, try to focus on the bigger picture of analyzing data and drawing appropriate conclusions.

Increases and decreases in motor vehicle fatalities can be monitored to help assess impacts of changes in policy. For example, in 1974, the US Federal government passed the National Maximum Speed Law, capping the speed limit on all interstate roads to 55mph. At the time, the goal was to improve fuel consumption in response to the 1973 oil embargo. However, when Congress repealed this law in 1995, allowing states to raise the legal speed limit, many wondered if that would also lead to an increase in accidents and fatalities. Frieman, Hedeker, and Richter (2009) took data from the Fatality Analysis Reporting System (FARS) to explore this issue, focusing in particular on the repeal of federal speed limit controls on road fatalities.

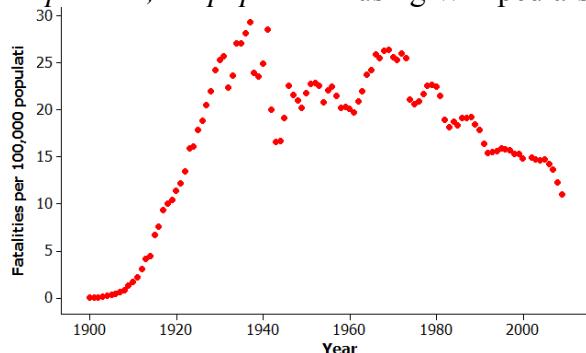
- (a) To the right is a portion of Wikipedia's "[List of motor vehicle deaths in the U.S. by year](#)" data (through 2013) on the number of total motor vehicle deaths in the United States per year. Notice the data are given in a *data table*: each row represents a different year and the second column is the number of deaths in the U.S. that year. Suggest a research question that you could investigate with the full dataset.

- (b) From the data provided here, do you see any trend or pattern to how the *number of deaths per year* is changing? Does this pattern make sense? What other information would you like to consider?

A *time plot* is a natural way to examine how a *variable* changes over time, but not about how other variables are also changing over the same time period. In particular, there are a lot more drivers on the roads now than in 1899, which is going to impact the number of deaths, so we really should take that into account. Typically, data like this would be scaled in some way, such as dividing by number of vehicle miles traveled by all cars that year or by the population size of the United States that year.

Year	Death
1899	26 <sup>[3]</sup>
1900	36
1901	54
1902	79
1903	117
1904	172
1905	252
1906	338
1907	581
1908	751
1909	1,174
1910	1,599
1911	2,043
1912	2,968
1913	4,079
1914	4,468
1915	6,779
1916	7,766
1917	9,630

Below is a [time plot](#) of *fatalities per 100,000 population* using Wikipedia's data from 1899 to 2013.



(c) Write a few sentences summarizing how this new variable was calculated and what you learn from the time plot of these data.

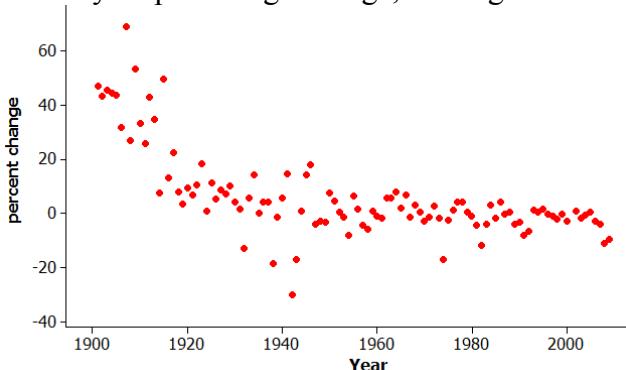
(d) In 1973 (before the speed limit was capped at 55 mph nationwide), the data report 25.507 fatalities per 100,000 people, compared to 21.134 in 1974 (after the national limit). Did the rate of fatalities decrease between these two years? If so, by how much? Does that appear to be a “large” difference? How are you deciding?

*Absolute differences*, like you calculated in (d), can be difficult to evaluate, especially if you don't consider the magnitude of values they come from. You might respond differently to a box of M&Ms coming up 4 candies short, than to a difference of 4 fatalities per 100,000 people (about 10,000 more deaths across the U.S. in one year). One way to take into account the magnitude of the data values you are comparing is to compute the *percentage change*:

$$\text{percentage change} = \frac{(\text{current rate} - \text{previous rate})}{\text{previous rate}} \times 100\%$$

(e) Calculate and interpret this value. In particular, is this a positive or negative value; what does that imply?

Below is a [time plot](#) of the year-to-year percentage change, starting in 1900.

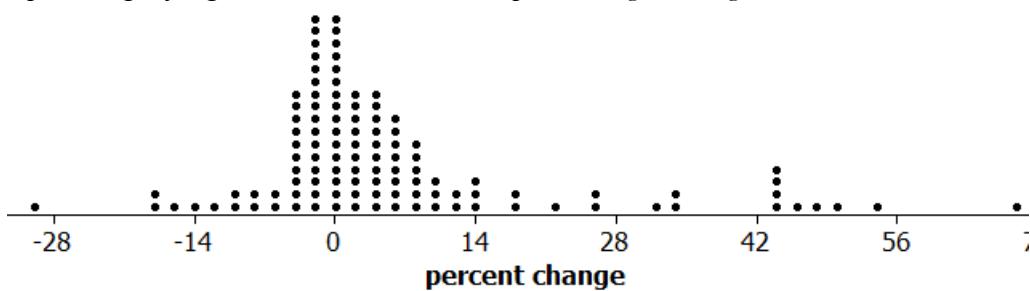


(f) Why doesn't this time plot start in 1899?

(g) Write a few sentences describing what you learn from this graph. Are there any unusual features? Can you conjecture an explanation for any of these features?

From this graph, we notice that there is *variability* in how these values change from year to year. Much of statistical investigation is measuring, accounting for, and trying to explain variability. In particular, we still can't judge whether the percentage change you computed between 1973 and 1974 is particularly large until we compare it to the percentage changes for the other years.

Below is a *dotplot* displaying the *distribution* of the *percentage change* values for this data set.



(h) Explain what each dot represents. What information is revealed by this plot that was less clear in the time plot? How might you use this graph to measure how unusual the 1973/1974 percentage change is? What information is lost in this graph compared to the time plot?

When looking at a distribution of a single variable like this, we are often interested in three key features:

- Center: What would you consider a “typical” value in the distribution?
- Variability: How clustered together or consistent are the observations?
- Shape: Are some values more common than others? Are the values symmetric about the center? Are there any unusual observations that don’t follow the overall pattern?

In describing these features, it is often helpful to summarize a characteristic with a single number. To summarize the center of the distribution, we often report the *mean* or the *median*. To summarize the variability of the distribution, we often report the *standard deviation* or the *interquartile range*. For now we will focus on the mean and standard deviation, a very common pairing.

**Definitions:** If there are  $n$  such values and we refer to them as  $x_1, x_2, \dots, x_n$ ,

- The *mean* is the average of all numerical values in your data set.
- The *median* is a value such that 50% of the data lies below and 50% of the data lies above that value
- The *standard deviation* measures the size of a “typical” deviation of the data values from the mean.

$$\text{mean } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

median position:  $(n+1)/2$

$$\text{standard deviation } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

**Quick Check:** Which set of values will have the largest standard deviation?

- (a) 1, 3, 5, 7, 9      (b) 1, 1, 5, 9, 9      (c) 5, 5, 5, 5, 5

(i) Open the [TrafficFatalitiesYears.txt](#) file and copy the *PercentChange* column into the [Descriptive Statistics](#) applet:

- Select and highlight all the data values in the text file and copy to the clipboard (e.g., ctrl-C or right click and select copy).
- In the applet, keep the **Includes header** box checked. (Note that the header must consist of just one word.)
- Press the **Clear** button and then click inside the Sample data window. Then paste the data from the clipboard (e.g., ctrl-V or right click and select paste).
- You can delete the \* (representing the percentage change for the initial year) or the applet will warn you that it is ignoring that value but allow you to proceed.
- Press **Use Data**.

What do you see?

Because the data set doesn’t consist of a lot of identical numerical values, the dotplot is not very informative, looking somewhat flat. We could round the data values to the nearest integer or we can “bin” the observations together a bit: group similar values into non-overlapping categories and then display the count of observations in each category. The dotplot shown before (h) is binned.

(j) In the applet, select the radio button for **Histogram** to have the applet bin the observations together and create a *histogram*. The height of each bar counts the number of observations that fall (uniquely) in each bin. How does the histogram representation compare to the original dotplot? Which display do you prefer? What is a potential downside to a histogram with not very many bins? [Hint: You can adjust the number of bins below the graph.]

(k) Now check the boxes for Actual in the Mean row and in the Std dev row to the left of the graph. Report these values and summarize what they tell you.

(l) We can use this information to help us measure how “large” the percentage change was after the repeal of the federal speed limit. Where does the percentage change you calculated in (e) fall in this distribution? (Give a pretty detailed description.) Would you consider it an atypical value?

(m) The percentage change between 1994 and 1995 (before and after the federal speed limit was repealed) was 1.74%. Where does that value fall in the above distribution? Does it appear unusually large?

Of course, one issue with this analysis is that after the repeal of the federal speed limit different states chose to implement changes to the speed limits in different years, so focusing on the 1994-95 change is too simplistic. One way to account for this *source of variability* is by focusing on an individual state. The [CAFatalities.xls](#) file contains the number of fatalities and population size (in thousands) for 1993-2012 from [www-fars.nhtsa.dot.gov/Trends/TrendsGeneral.aspx](http://www-fars.nhtsa.dot.gov/Trends/TrendsGeneral.aspx).

(n) What was (roughly) the California population in 1993?

(o) Use Excel to compute the fatality rates per thousand residents (e.g., In cell E2, enter =C2/D2, then put the cursor back over that cell and double click on the lower right corner to “fill down” the column. Remember to name the column!). What was the fatality rate in 1993? (Include “measurement units” with your answer.)

(p) Now compute the percentage change in fatality rates (e.g., In cell F3, enter = (E3-E2) / E2 \* 100 and fill down). What was the percentage change from 1993 to 1994? Was it positive or negative?

(q) Use technology to create a dotplot or histogram of the *percentage change in fatality rates* for California during this time frame, and calculate the mean and standard deviation. Compare these results to what you found above for the entire United States.

(r) In particular, is the standard deviation in (q) larger or smaller than for the U.S. data? Give some explanations why the standard deviation has changed this way. [Hint: Focus on issues other than the sample size (the number of years of data), because the number of observations is *not* generally an indicator of how large the “typical deviation from the mean” will be. In fact, if we consider the smaller population size in CA, we might expect there to be *more* variability in the rate of accidents per year...]

(s) On Jan. 7, 1996, California allowed rural speed limits to increase from 65 to 70 mph and urban roads to increase from 55 to 65 mph. What is the percentage change in the fatality rate for California between 1995 and 1996? Is this a particularly large value in the distribution for California? [Be sure to justify your answer.]

(t) Suppose we did find a large increase or decrease in a state’s fatality rate before and after the change in speed limit laws. Would it be reasonable to conclude that the change in speed limit laws *caused* the change in fatality rate? Why or why not? What else should we consider?

### Study Conclusions

Many researchers are interested in whether changes in speed limits have had an impact on the number of traffic fatalities. In this investigation, we focused on the number of fatalities per 100,000 in the United States. (You could also look at other rates like fatalities per 100 million miles traveled.) We see a decreasing trend in the fatality rate over time, though the rates appear to be leveling off after about 1950. The second largest decrease (more than one standard deviation from the mean) occurred in 1976, after the imposition of a federal maximum speed limit. We did not see a similar increase when the federal maximum speed limit was repealed, though not all the states changed their speed limits at this time. A follow-up study would examine whether the fatality rates differed between states that did and did not increase speed limits (see Example 3.2).

**Discussion:** Statistical analysis often involves looking for patterns in data. In the age of the “data deluge,” this is becoming increasingly important. This investigation should have raised your awareness of several key issues that will permeate the studies you analyze in this text.

- It's important to determine which variables are most relevant to the research question and whether you can collect the data you need to answer the question
- Simple graphs can be very informative, but you should also take care in considering the most meaningful variable representation of what you are studying even before you begin graphing.
- It is often misleading to view results in isolation. The change in traffic fatalities needs to be considered in the context of other years and other changes.
- It is imperative to consider variability. The percentage change will fluctuate from year to year naturally, and we should not make large-scale conclusions of the impact of policy changes without considering the size of a change with respect to the natural variability. *Standardizing* is one way to interpret an observation's relative position in a distribution in terms of “how many standard deviations is the observation from the mean” =  $(\text{observation} - \text{mean}) / \text{standard deviation}$ .
- It is very difficult to isolate the cause of a change. There easily could have been other reasons for a decrease in traffic deaths between two years — even just awareness of the issue could have changed drivers' behavior.

A key skill in statistical analysis is communicating the results of your analysis to different audiences, including those with little experience with quantitative data. With each investigation we will present a Study Conclusions box as above as a model for an effective summary of the analysis of the research question, which often includes statements of new questions.

With most investigations we will also provide a follow-up practice problem for you to try on your own to assess your understanding of the material.

### Practice Problem A

Go to <http://www-fars.nhtsa.dot.gov/> and

- Select the FARS Data Tables tab (it should be orange).
- Press the Trends tab (so it turns orange) underneath.
- In the next line of tabs, select General.
- Use the pull-down menu to select a state other than California.
- Scroll down to find the *Fatalities and Fatality Rates* table. Then click on the Export to export the data to a text or Excel file.
  - (a) Select the *Fatality Rate per 100,000 Population* column (remember to use one word header names) and use technology (e.g., the [Descriptive Statistics](#) applet) to create a time plot (e.g., select the radio button under the graph) and a dotplot. Write a few sentences summarizing the results.
  - (b) Search online to determine whether the state you chose changed the maximum speed limit(s) after the 1995 repeal and if so, in what year. Do you see a large percentage change after that year? (Remember you have the individual data files too.)
  - (c) How do the other rate variables (per motor vehicles or per miles driven) behave? Much differently than per 100,000 people?
  - (d) When did your state enact seat belt laws? Does that appear to correspond to a change in fatality rate?

## Technology Detour – Scraping Data in R

In R you can “scrape” the data table from the webpage directly. Here is one approach:

Specify the URL of the webpage

```
> myurl =
"https://en.wikipedia.org/wiki/List_of_motor_vehicle_deaths_in_U.S._by_year"
```

Because there are two tables on this page, we want to only take the first table, and we are telling R the type of data in each column. Note: R is *very* sensitive to capitalization.

```
> library(XML); library(RCurl)
> mypage = getURL(myurl, .opts=list(ssl.verifyPeer = FALSE))
> table1 = readHTMLTable(mypage, which=1, colClasses = c("integer", "numeric",
"numeric", "numeric", "character", "numeric", "character"))
```

Let's get a sense of what we have created

```
> head(table1)
```

We notice that there are missing values (NA), including some of the years when they had footnotes. There are also the commas in the population numbers and the percentage sign actually pose problems to using these data as numbers. We also notice that the column names are very long, so we can change those.

```
> names(table1) = c("year", "death", "vmt", "fatalitiesVMT", "population",
"fatalities", "percentchange")
```

So first, let's get rid of the commas in the population results and tell R to treat this data as numeric.

```
> table1$population = as.numeric(gsub(",", "", table1$population))
```

We can do the same trick with the percentage change data

```
> table1$percentchange = as.numeric(gsub("%", "", table1$percentchange))
```

(In fact, there is a similar trick for getting rid of the footnotes before converting those columns.)

Finally we can create the time plot. Because of the missing values, we will create an object that consists of the two variables first.

```
> plot(table1$year, table1$percentchange)
```

If you wanted only the percentage change data, you can use

```
> stripchart(table1$percentchange)
```

But this graph can be hard to read. One of our goals in this course will be to look at better types of graphs. Although there are different ways to use the technology, it's also important to think about broader goals of how we represent and communicate data.

For many software packages, you can mouse over the table (maybe leaving out the column names at first) and copy and paste directly into a spreadsheet. But you can still expect several “data cleaning” steps (e.g., how to handle the footnotes) along the way. On this webpage, table 2 is much easier to copy and paste than table 1, for example.

### Investigation B: Random Babies

In the previous investigation, you looked at historical data. In this investigation, the goal is to explore a *random process*. We apologize in advance for the absurd but memorable process below. Another example of a random process is coin flipping.

Suppose that on one night at a certain hospital, four mothers give birth to baby boys. As a very sick joke, the hospital staff decides to return babies to their mothers completely at random. Our goal is to look for the pattern of outcomes from this process, with regard to the issue of how many mothers get the correct baby. This enables us to investigate the underlying properties of this child-returning process, such as the probability that at least one mother receives her own baby.

- (a) *Prediction:* If we were to repeat this process over many many nights, predict how often (what percentage of the nights) at least one mother will receive her own baby. Predict how often all four mothers will receive her own baby.

Because it is clearly not feasible to actually carry out this horrible child-returning process over and over again, we will instead *simulate* the random process to investigate what would happen in the long run.

Suppose the four boys were named Marvin Miller, Willie Williams, Billy Brown, and Sam Smith. Take four index cards and write the first name of each boy on a different card. Now take a sheet of paper and divide it into four regions, one for each mom. Next shuffle the index cards, face down, and randomly deal them back to the mothers. Flip the cards over and count the number of moms who received her own baby.

- (b) How many mothers received the correct baby: 0      1      2      3      4

- (c) Did everyone in class obtain the same number of matches? If not, why not?

- (d) Repeat your shuffling process 4 more times (for a total of 5 *trials*):

<b>Trials</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>Number of matches</b>					

Then pool your results with those of your classmates. Tally how many outcomes resulted in each value, and also compute the proportion of the trials that resulted in each value.

<b>Number of matches</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>Count</b>					
<b>Proportion</b>					

- (e) Calculate the proportion of trials for which at least one mother received her own baby [*Hint:* Note, there are two ways you can find this value: Sum up all the proportions corresponding to 1, 2, 3, or 4 matches or consider the “complement” = no matches and take one minus that proportion.]

(f) The proportion you have calculated is an estimate of how often this “event” happens in the long run. How does this proportion compare to your prediction in (a)?

(g) How could we improve our estimate of how often this event happens?

### Simulation Analysis

Open the [Random Babies](#) applet and press **Randomize**.

Notice that the applet randomly returns babies to mothers and determines how many babies are returned to the correct home (by matching diaper colors). The applet also counts and graphs the resulting number of matches. Uncheck the

**Animate** box and press **Randomize** a few more times. You should see the results accumulate in the table and the histogram. Click on the bar representing the outcome of zero mothers receiving the correct baby. This shows you a “time plot” of the proportion of trials with 0 matches vs. the number of trials. Set the **Number of trials** to 100 and press

**Randomize** a few times, noticing how the behavior of this graph changes.



**Animate**

Number of babies:

Number of trials:

(h) Describe what you learn about how the (cumulative) proportion of trials that result in zero matches changes as you simulate more and more trials of this process.

**Definition:** A [random process](#) generates observations according to a random mechanism, like a coin toss. Whereas we can't predict each individual outcome with certainty, we do expect to see a long-run pattern to the results. The [probability](#) of a random *event* occurring is the **long-run proportion** (or relative frequency) of times that the event would occur if the random process were repeated over and over under identical conditions. You can *approximate* a probability by simulating (i.e., artificially recreating) the process many times. Simulation leads to an empirical estimate of the probability, which is the proportion of times that the event occurs in the simulated repetitions of the random process. Increasing the number of repetitions generally results in more accurate estimates of the long-run probabilities.

(i) After at least 1000 trials, complete the table below.

Number of matches	0	1	2	3	4
Proportion					

(j) Based on your answer to (i), what is your estimate for the probability that at least one mother receives her own baby? Do all of your classmates obtain the same estimate?

(k) Consider the table of Cumulative Results in the applet. One value for the *number of matches* is fairly unlikely, but does occur once in a while. Which outcome is this?

(l) One outcome is actually impossible to occur. Which outcome is this? Explain why it is impossible.

(m) Calculate the average number of matches for your 1000 (or however many you performed) trials.  
[Hint: Use a *weighted average*  $(\sum x_i f_i)/N$  where  $x_i$  is the number of matches and  $f_i$  is the corresponding frequency, and  $N$  is the number of repetitions you simulated.]

(n) Select the **Average** radio button above the time plot. Describe the behavior of this graph and comment on what it reveals about the behavior of the average number of matches over many trials.

### Exact Mathematical Analysis

One disadvantage to using simulation to estimate a probability like this is that everyone will potentially obtain a different estimate. Even with a very large number of trials, your result will still only be an estimate of the actual long-run probability. For this particular scenario however, we can determine exact theoretical probabilities.

First, let's list out all possible outcomes for returning four babies to their mothers at random. We can organize our work by letting 1234 represent the outcome where the first baby went to the first mother, the second baby to the second mother, the third baby to the third mother, and the fourth baby to the fourth mother. In this scenario, all four mothers get the correct baby. As another example, 1243 means that the first two mothers get the right baby, but the third and fourth mothers have their babies switched.

**Definition:** A sample space is a list of all possible outcomes of a random process.

All of the possible outcomes are listed below:

**Sample Space:**

1234	1243	1324	1342	1423	1432
2134	2143	2314	2341	2413	2431
3124	3142	3214	3241	3412	3421
4123	4132	4213	4231	4312	4321

In this case, returning the babies to the mothers completely at random implies that the outcomes in our sample space are equally likely to occur (*outcome probability* = 1 / *number of possible outcomes*).

- (o) How many different outcomes are there for returning four babies to their mothers? What is the each outcome's probability of occurring for any trial?

You could have determined the number of possible outcomes without having to list them first. For the first mother to receive a baby, she could receive any one of the four babies. Then there are three babies to choose from in giving a baby to the second mother. The third mother receives one of the two remaining babies and then the last baby goes to the fourth mother. Because the number of possibilities at one stage of this process does not depend on the outcome (which baby) of earlier stages, the total number of possibilities is the product  $4 \cdot 3 \cdot 2 \cdot 1 = 24$ . This is also known as  $4!$ , read “4 factorial.” Because the above outcomes are equally likely, the probability of any one of the above outcomes occurring is  $1/24$ . Although these 24 outcomes are equally likely, we were more interested above in the probability of 0 matches, 1 match, etc.

**Definition:** A [random variable](#) maps each possible outcome of the random process (the sample space) to a numerical value. We can then talk about the *probability distribution* of the random variable. These random variables are usually denoted by capital roman letters, e.g.,  $X$ ,  $Y$ . A random variable is *discrete* if you can list each individual value that can be observed for the random variable.

- (p) For each of the above outcomes, indicate (next to the outcome above) how many mothers get the correct baby. [Hint: For outcome 1243, mothers one and two receive their own babies so the value of the number of correct matches is 2.]

- (q) In how many of the outcomes did zero mothers receive the correct baby?

**Probability Rule:** When the outcomes in the sample space are equally likely, the probability of any one of a set of outcomes (an event) occurring is the number of outcomes in that set divided by the total number of outcomes in the sample space.

- (r) To calculate the exact probability of 0 matches, divide the number of outcomes with 0 matches by the total number of possible outcomes. How does this result compare to your estimate from the simulation?

$$P(X = 0) =$$

Comparison:

(s) Use this method to determine the exact probabilities for each possible value for the number of matches. Express these probabilities as fractions and as decimals in the table below.

**Probability Distribution for # of correct matches:**

# matches	0	1	2	3	4
Probability (fraction)					
Probability (decimal)					

How do these theoretical probabilities compare to the empirical estimates you found in the simulation (question (i))?

(t) What is the sum of these five probabilities? Why does this make sense?

(u) What is the probability that *at least one* mother receives the correct baby? [Hint: Determine this two different ways: first by adding the probabilities of the corresponding values, and then by taking one minus the probability that this event does not happen.] How does this compare to the simulation results?

**Probability rules:**

- The sum of the probabilities for all possible outcomes equals one.
- Complement rule: The probability of an event happening is one minus the probability of the event not happening.
- Addition rule for disjoint events: The probability of at least one of several events is the sum of the probabilities of those events *as long as* there are no outcomes in common across the events (i.e., the events are *mutually exclusive* or *disjoint*).

We can also consider the expected value of the number of matches, which is interpreted as the long-run average value of the random variable. For a discrete random variable we can calculate the expected value of the random variable X by again employing the idea of a weighted average of the different possible values of the random variable, but now the “weights” will be given by the probabilities of those values:

$$E(X) = \sum_{\text{all possible values}} (\text{value}) \times (\text{probability of value})$$

(v) Calculate the expected value of the number of matches. Comment on how it compares to the average value you obtained in the simulation.

(w) Is the expected value for the number of matches equal to the most probable outcome? If not, explain what is meant by an “expected” value.

Notice that if we wanted to compute the average number of matches, say after 1000 trials, we would look at a weighted average:

$$\bar{x} = \sum \frac{1+0+1+2+0+\dots}{1000} = \sum \frac{(\# \text{of } 0s) \times 0 + (\# \text{of } 1s) \times 1 + (\# \text{of } 2s) \times 2 + (\# \text{of } 4s) \times 4}{1000}$$

But from the results we saw above, each term (# of)/1000 terms converges to the probability of that outcome as we increase the number of repetitions, giving us the above formula for  $E(X)$ . So we will interpret the expected value as the long-run average of the outcomes.

Another property of a random variable is its variance. This measures how variable the values of the random variable will be. For a discrete random variable we can again use a type of weighted average, based on the probabilities of each value and the squared distances between the possible values of the random variable and the expected value.

$$V(X) = \sum_{\text{all possible values}} (\text{value} - E(X))^2 \times (\text{probability of value})$$

(x) Calculate the variance of the number of matches. Also take the square root to calculate the standard deviation  $SD(X)$ .

$$V(X) = \quad \quad \quad SD(X) =$$

We will interpret this standard deviation similarly to how we did in Investigation A: how far the outcomes tend to be from the expected value. Here we are talking in terms of the probability model; in Investigation A we were talking in terms of the historical data.

(y) Confirm that the value of the standard deviation you calculating makes sense considering the possible outcomes for the random variable.

**Discussion:** Notice that we have used two methods to answer questions about this random process:

- *Simulation* – running the process under identical conditions a large number of times and seeing how often different outcomes occur
- *Exact mathematical calculations* using basic rules of probability and counting

This approach of looking at the analysis using both simulation and exact approaches will be a theme in this course. We will also consider some approximate mathematical models as well. You should consider these multiple approaches as a way to assess the appropriateness of each method. You should also be aware of situations where one method may be preferable to another and why.

**Practice Problem B.A**

Suppose three executives (Annette, Barb, and Carlos) drop their cell phones in an elevator and blindly pick them back up at random.

- (a) Write out the sample space using *ABC* notation for the outcomes.
- (b) Carry out the exact analysis to determine how the probability of at least one mother receiving her own baby.
- (c) Calculate the expected number of matches for 3 mothers. How does this compare to the case with 4 mothers?
- (d) Use the Random Babies applet to check your results.

**Practice Problem B.B**

Reconsider the Random Babies. Now suppose there were 8 mothers involved in this random process.

- (a) Calculate the (exact) probability that all 8 mothers receive the correct baby. [Hint: First determine how many possible outcomes there are for returning 8 babies to their mothers.]
- (b) Calculate the probability that exactly 7 mothers receive the correct baby.
- (c) Using the Random Babies applet, approximate the probability that at least one of the 8 mothers receives the correct baby. How does your approximation compare to the probability of this event with 4 mothers?
- (d) Using the Random Babies applet, approximate the expected value for how many of the eight mothers will receive the correct baby. How does your approximation compare to the situation with 4 mothers?

## CHAPTER 1: ANALYZING ONE CATEGORICAL VARIABLE

In this chapter, you will begin to analyze results from statistical studies and focus on the process of statistical inference. In particular, you will learn how to assess evidence against a particular claim about a random process.

### Section 1: Analyzing a process probability

Investigation 1.1: Friend or foe – Inference for a proportion

Probability Exploration: Mathematical Model

Probability Detour: Binomial Random Variables

Investigation 1.2: Do names match faces – Bar graph, hypotheses, binomial test (technology)

Investigation 1.3: Heart transplant mortality – Factors affecting p-value

Investigation 1.4: Kissing the right way – Two-sided p-values

Investigation 1.5: Kissing the right way (cont.) – Interval of plausible values

Investigation 1.6: Improved baseball player – Types of error and power

Probability Exploration: Exact Binomial Power Calculations

### Section 2: Normal approximations for sample proportions

Investigation 1.7: Reese's pieces – Normal model, Central Limit Theorem

Probability Detour: Normal Random Variables

Investigation 1.8: Is ESP real? – Normal probabilities,  $z$ -score

Investigation 1.9: Halloween treat choices – Test statistic, continuity correction

Investigation 1.10: Kissing the right way (cont.) –  $z$ -interval, confidence level

Investigation 1.11: Heart transplant mortality (cont.) – Plus Four/Adjusted Wald

### Section 3: Sampling from a finite population

Investigation 1.12: Sampling words – Biased and random sampling

Investigation 1.13: *Literary Digest* – Issues in sampling

Investigation 1.14: Sampling words (cont.) – Central Limit Theorem for  $\hat{p}$

Investigation 1.15: Freshmen Voting Patterns – Nonsampling errors, hypergeometric distribution

Probability Detour: Hypergeometric Random Variables

Investigation 1.16: Teen hearing loss – One sample  $z$ -procedures

Investigation 1.17: Cat households – Practical significance

Investigation 1.18: Female senators – Cautions in inference

Example 1.1: Predicting Elections from Faces

Example 1.2: Cola Discrimination

Example 1.3: Seat Belt Usage

Appendix: Stratified random sampling

## SECTION 1: ANALYZING A PROCESS PROBABILITY

In this investigation you will be introduced to a basic statistical investigation as well as some ideas and terminology that you will utilize throughout the course. You will combine ideas from the preliminary investigations: examining distributions of data and simulating models of random processes to help judge how unusual an observation would be for a particular probability model.

### Investigation 1.1: Friend or Foe?

In a study reported in the November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction (Hamlin, Wynn, and Bloom, 2007). In other words, do children who aren't even yet talking still form impressions as to someone's friendliness based on their actions? In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "googly" eyes glued onto it) that could not make it up a hill in two tries. Then the infants were shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character (the "helper" toy) and one where the climber was pushed back down the hill by another character (the "hinderer" toy). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer characters) and asked to pick one to play with. Videos demonstrating this component of the study can be found at <http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html>.

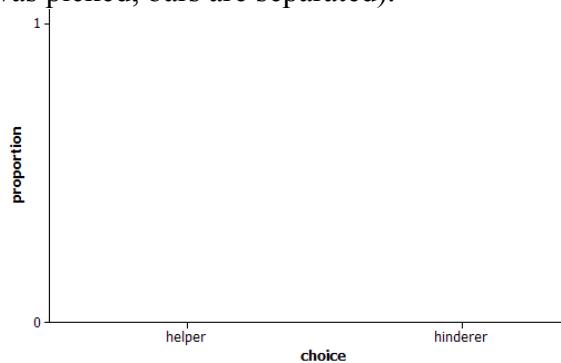
One important design consideration to keep in mind is that in order to equalize potential influencing factors such as shape, color, and position, the researchers varied the colors and shapes of the wooden characters and even on which side the toys were presented to the infants. The researchers found that 14 of the 16 infants chose the helper over the hinderer.

### Summarizing the Observed Data

(a) What proportion of these infants chose the helper toy? Is this more than half (a majority)? Also sketch by hand a simple bar graph to display the results for this sample (one bar for each toy, with heights representing the proportion of times that toy was picked, bars are separated).

Proportion?

Majority?



### Drawing Conclusions Beyond the Data

(b) Suggest some explanations for why this majority might have occurred. [Hint: Why do you think the researchers carried out this study?]

We can place these explanations into two main categories:

- (1) There is something to the theory that infants are genuinely more likely to pick the helper toy (for some reason).
- (2) Infants choose randomly and we happened to get “lucky” and find most infants picking the helper toy in our study.

(c) Why are we not considering something like color of the helper toy as a possible explanation?

(d) So for the two explanations we are still considering, how might you choose between them? In particular, how might you convince someone that option (2) is *not* plausible for this study?

**Discussion:** As you saw with the Random Babies, we can simulate the outcomes of a random process to help us determine which types of outcomes are more or less likely to occur. In this case, we can easily simulate what “could have happened” if the infants are randomly guessing. In other words, our analysis is going to assume the researchers’ conjecture is wrong (for the moment) and that infants really are just blindly picking one toy or the other without any regard for whether it was the helper toy or the hinderer.

(e) Suggest a method for carrying out this simulation of infants picking equally between the two toys.

(f) Explain why looking at such “could have been” results will be useful to us.

**Discussion:** We will call the assumption that these infants have no genuine preference between the toys the null model. Performing lots of repetitions under this model will enable us to see the pattern of results (number who choose the helper toy) *when we know the infants have no preference*. Examining this null distribution will in turn help us to determine how unusual it is to get 14 infants picking the helper toy where there is no genuine preference. If the answer is that the result observed by the researchers (14 of 16 choosing the helper toy) would be very surprising for infants who had no real preference, then we would have strong evidence to conclude that infants really *do* prefer the helper. Why? Because otherwise, we would have to believe a very rare coincidence just happened to occur in this study.

**Summing up: An observed outcome that would rarely happen if a claim were true provides strong evidence that the claim is not true.**

## Simulation

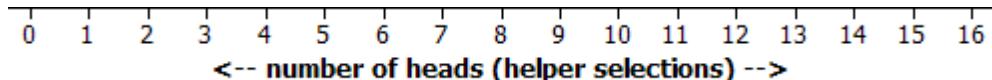
(g) Flip a coin 16 times, representing the 16 infants in the study (one *trial* or *repetition* from this random process). Let a result of heads mean that the infant chooses the helper toy, tails for the hinderer toy. Tally the results below and count how many of the 16 chose the helper toy:

“Could have been” distribution	
Heads (helper toy)	Tails (hinderer toy)
Total number of heads in 16 tosses:	

(h) Repeat this two more times. Keep track of how many infants, out of the 16, choose the helper. Record this number for all three of your repetitions (including the one from the previous question):

Repetition #	1	2	3
Number of (simulated) infants who chose helper			

(i) Combine your simulation results for each repetition with your classmates on the scale below. Create a *dotplot* by placing a dot above the numerical result found by each person.



(j) Did everyone get the same number of heads every time? What is an average or typical number of heads? Is this what you expected? Explain.

(k) How is it looking so far – is the actual study result (14 picking the helper toy) consistent with the outcomes that could have happened under the null model? Which explanation (1) or (2) do you think is more plausible based on this null distribution? Explain your reasoning.

We really need to simulate this hypothetical random selection process hundreds, preferably thousands of times. This would be very tedious and time-consuming with coins, so let's turn to technology.

(l) Use the [One Proportion Inference](#) applet to simulate these 16 infants making this helper/hinderer choice, still assuming the null model that infants have no real preference and so are equally likely to choose either toy.

- Keep the **Probability of heads** set to 0.5.
- Set the **Number of Tosses** to 16.
- Keep the **Number of repetitions** at 1 for now.
- Press **Draw Samples**.

Probability of heads:	<input type="text" value="0.5"/>
Number of tosses:	<input type="text" value="16"/>
Number of repetitions:	<input type="text" value="1"/>
<input checked="" type="checkbox"/> Animate	
<b>Draw Samples</b>	

Report the number of heads (i.e., the number of infants who choose the helper toy) for this “could have been” (under the null model) outcome.

(m) **Uncheck** the **Animate** box and press **Draw Samples** four more times, each time recording the number of the 16 infants who choose the helper toy. Did you get the same number of heads all five times?

(n) Now change the **Number of repetitions** to **995** and press **Draw Samples**, to produce a total of 1000 repetitions of this null process of tossing a coin 16 times. Comment on the “null distribution” of the number of infants who choose the helper toy, across these 1000 repetitions. In particular, comment on where this distribution is centered, on how spread out or variable it is (e.g., min and max values or standard deviation), and on the distribution’s general shape.

Center:

Variability:

Shape:

### Measuring “Rareness”

As in Investigation A, we want a method for measuring how unusual our observation (in this case 14) is in the distribution (in this case the null distribution). We could just see whether  $P(X = 14)$  is small. But if we increase the sample size (e.g., 160 infants), than *any* one particular outcome will have a small probability. So we want to judge how extreme the observation is relative to the other observations in the distribution. One way to do this is to count how many observations are even more extreme. For example, if we tell you that only 1% of rattlesnakes are longer than 2.5 meters, then you know to be very surprised to see a 3 meter rattlesnake and you may even begin to think that what you are looking at is not a rattlesnake at all! Such a judgement works for any distribution and does not depend on knowing specific characteristics like the mean and standard deviation in order to be meaningful.

(o) Report how many and what proportion of these 1000 samples produced 14 or more infants choosing the helper toy:

- Enter **14** in the **As extreme as** box
- Press the **Count** button.

Number of repetitions:  
Proportion of repetitions:

(p) Did everyone in your class get the same proportion? Does this surprise you?

(q) Based on the proportion you found in (o), would you say that the actual result obtained by the researchers is very rare, somewhat rare, or not very rare, *under the null model that infants have no preference* and so choose blindly? Circle one of these three:

Very rare

Somewhat rare

Not very rare

(r) So, based on your simulation results, would you say that the researchers have very strong evidence that these infants' selection process is not like flipping a coin, and instead the more plausible explanation is that these infants do have a preference for the helper toy?

### Terminology Detour

The long-run proportion of times that an event happens when its random process is repeated indefinitely is called the probability of the event. We can approximate a probability empirically by simulating the random process a large number of times and determining the proportion of times that the event happens.

More specifically, the probability that the random process specified by the null model would produce results as or more extreme as the actual study result is called a p-value. Our analysis above approximated this p-value by simulating the infants' random selection process a large number of times and finding how often we obtained results at least as extreme as the actual data. You can obtain better and better approximations of this p-value by using more and more repetitions in your simulation.

A small p-value indicates that the observed data would be surprising to occur through the random process alone, if the null model were true. Such a result is said to be statistically significant, providing evidence against the null model (so we don't believe the discrepancy arose just by chance but instead reflects a genuine tendency). The smaller the p-value, the stronger the evidence against the null model. There are no hard-and-fast cut-off values for gauging the smallness of a p-value, but generally speaking:

- A p-value above 0.10 constitutes *little or no evidence* against the null model.
- A p-value below 0.10 but above 0.05 constitutes *moderate evidence* against the null model.
- A p-value below 0.05 but above 0.01 constitutes *strong evidence* against the null model.
- A p-value below 0.01 constitutes *very strong evidence* against the null model.

## Scope of Conclusions

What bottom line does our analysis lead to? Do infants in general show a genuine preference for the “helper” toy over the “hinderer” one? Well, there are rarely definitive answers when working with real data, but our analysis reveals that the study provides strong evidence that these infants are not behaving as if they were tossing coins. That is, we have strong evidence that the probability of choosing the helper toy in this process is greater than 0.5. Because the researchers controlled for other possible explanations for the observed preference, we will conclude that there is convincing evidence that infants really do have a preference.

(s) Can we reasonably claim that it was the actions of the objects in the videos that influenced the infants’ choices or is “random luck” still a *plausible* (believable) explanation? Explain.

(t) Can we reasonably extend this conclusion beyond the 16 infants in the study? Explain. [Note: We will discuss this issue in more detail later. What does your intuition tell you for now?]

## Study Conclusions

If we assume the infants choose between the two toys with equal probability, than we can model this random process with coin tossing. Approximating the *p-value*, we found that getting 14 or more infants choosing the helper toy, if the infants are choosing at random, is very unlikely by chance alone. This means that the data provide very strong statistical evidence to reject this “no preference” model and conclude that these infants’ choices are actually governed by a process where there is a genuine preference for the helper toy (or at least that it’s more complicated than each infant flipping a coin to decide). Of course, the researchers really care about whether infants *in general* (not just the 16 in this study) have such a preference. Extending the results to a larger group of infants depends on whether it’s reasonable to believe that the infants in this study are *representative* of a larger group of infants on this question. We cannot make this judgment without knowing more about how these infants were selected. However, you may be able to argue that these 16 infants’ choices are representative of the larger process of viewing the videos and selecting a toy to play with.

## Summary

Let's take a step back and consider the reasoning process and analysis strategy that we have employed here. Our reasoning process has been to start by supposing that infants in general have no genuine preference between the two toys (our null model), and then ask what results we expect to see under this null model (the null distribution) and whether the results observed by the researchers would be unlikely to have occurred just by random chance assuming this null model. To recap:

- We summarize the study results with one number, called a [statistic](#).
- Assume the null model is true, and simulate the random process under this model, producing data that “could have been” seen in the study if the null model were true. Calculate the value of the statistic from these “could have been” data. Then repeat this simulation a large number of times, generating the “null distribution” of the values of the statistic.
- Evaluate the strength of evidence against the null model by considering how extreme the observed value of the statistic is in the null distribution. If the observed value of the statistic falls in the tail of the null distribution, then *reject* the null model as not plausible. Otherwise, consider the null model to be [plausible](#) (but not necessarily true, because other models might also be plausible).

In this study, our statistic is the number (or the proportion) of the 16 infants who chose the helper toy. We assume that infants do not prefer either toy (the null model) and simulate the hypothetical random selection process a large number of times under this assumption. We noted that our actual statistic (14 of 16 choosing the helper toy) is in the upper tail of the simulated null distribution. Such a “tail result” indicates that the data observed by the researchers would be very surprising if the null model were true, giving us strong evidence against the null model. So instead of thinking the researchers just got that lucky that day, a more reasonable conclusion would be to “reject that null model.” Therefore, this study provides very strong evidence to conclude that these infants really do prefer the helper toy and were not essentially flipping a coin in making their selections.

**Important note:** If you do not obtain a small p-value, all you can conclude is that the null model is *plausible* for the overall process, but that other values could be plausible for the long-run probability as well. You have in no way *proven* that the hypothesized value is correct, just that you don't have convincing evidence it is false. In such a case, statisticians are very careful to say they *fail to reject* the null hypothesis, rather than saying anything that conveys they “accept” the null hypothesis.

## Practice Problem 1.1A

- (a) In a second experiment, the same events were repeated but the object climbing the hill no longer had the googly eyes attached. The researchers wanted to see whether the preference was made based on a social evaluation more than a perceptual preference. Suppose they found ten of sixteen 10-month-olds chose the helper in this study. If you were to use a coin to carry out a simulation analysis to evaluate these results: how many times would you flip the coin? How many times would you repeat the process? You would then determine the proportion that are (*choose one*:  $\leq, \geq, \neq$ ) (*choose one*: 8, 10, 14, 16, 1000).
- (b) Use the [One Proportion Inference](#) applet to approximate the p-value and compare to the p-value of the original study. Explain why this comparison makes intuitive sense. [Be sure to explain what values you input into the applet.]
- (c) What does the statistical significance of the first study and not of the second tell the researchers?

## Probability Exploration: Mathematical Model

(a) Did everyone in your class obtain the same approximate p-value from their simulation of 1000 repetitions? If not, are these approximate p-values all fairly close to each other?

Simulating 1000 repetitions is generally good enough to produce a reasonable approximation for the exact p-value. But in this case you can also determine this probability (p-value) exactly (as if the process were repeated infinitely often) using what are called binomial probabilities. The probability of obtaining  $k$  successes in a sequence of  $n$  trials with success probability  $\pi$  on each trial is:

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \text{ where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Here  $X$  represents a random variable which counts the number of successes in  $n$  trials. The short-hand notation for saying  $X$  follows a binomial distribution is  $X \sim \text{Bin}(n, \pi)$ . (See Probability Detour below.)

Use this expression to determine the exact probability of obtaining 14 or more successes (infants who choose the helper toy) in a sequence of 16 trials, under the null model that the underlying success probability on each trial,  $\pi$ , is 0.5:

(b) First determine the probability of exactly 14 successes, so  $k = 14$ ,  $n = 16$ , and  $\pi = 0.5$ :

$$P(X = 14) =$$

(c) Repeat (b) to find the (exact) probability of obtaining 15 successes. Then repeat to find the probability of obtaining 16 successes.

$$P(X = 15) =$$

$$P(X = 16) =$$

(d) Add these three probabilities to determine the (exact) p-value for 14 or more successes. Did the approximate p-value from your simulation come close to this value? Which value is “better”?

$$P(X \geq 14) =$$

Comparison:

The exact p-value (to four decimal places, see Technology Detour) turns out to be 0.0021. We can interpret this by saying that *if* infants really had no preference and so were randomly choosing equally between the two toys, there’s only about a 0.21% chance that 14 or more of the 16 infants would have chosen the helper toy. [Or, similar to what we saw in the simulation, such a result would only happen roughly twice in 1,000 samples.] Because this probability is quite small, the researchers’ data provide very strong evidence that having 14 of the infants pick the helper toy did not occur by chance alone and therefore strong evidence that infants in general really *do* have a preference for the nice (helper) toy.

## Probability Detour – Binomial Random Variables

A *binomial random variable* counts the number of successes in a random process with the following properties:

- Each trial results in “success” or “failure.”
- The trials are independent: The outcome of one trial does not change the probability of success on the next trial.
- The probability of success,  $\pi$ , is constant across the trials.
- There are a fixed number of trials,  $n$ .

If  $X$  counts the number of successes, then we can say  $X \sim \text{Binomial}(n, \pi)$ .

In this case, the probability calculation follows from several simple probability rules:

- When the trials are independent (the outcome on one trial does not affect the probability of success on another trial), the probability of several outcomes occurring simultaneously is the product of the probabilities of the individual outcomes.
- The probability of *one of* several mutually exclusive events (events that can't occur simultaneously, e.g., rolling a sum of six with two dice or rolling a sum of seven) occurring is the sum of the probabilities of the individual outcomes. By mutually exclusive, we mean the events do not share any outcomes (e.g., we can't roll both a six and a seven in the same roll).

How do these rules help us? Well, one of the probabilities we want to calculate is that among the 16 infants, 14 will choose the helper toy and 2 choose the hinderer toy. When the null model is true, meaning that infants have no preference, this amounts to finding the probability of getting exactly 14 heads in 16 coin tosses. One such sequence of coin toss outcomes with 14 heads and 2 tails is:

HTHHHHHHHHHTHHHHH

The first rule tells us that the probability of this sequence is

$$P(\text{HTHHHHHHHHHTHHHHH}) = P(H)^{14} \times P(T)^2$$

But this is not the only sequence that gives us 14 heads. We could have had

TTHHHHHHHHHHHHHH

which has the same probability of occurring as the first sequence above. The second rule tells us that the probability of *either* of these sequences occurring is simply the sum of their individual probabilities. So then the question is how many different sequences correspond to 14 heads and 2 tails? This is where the

*binomial coefficient* comes in handy. This binomial coefficient, denoted by  $C(n, k)$  or  $\binom{n}{k}$ , counts the

number of ways there are to select  $k$  items from a set of  $n$  items. In this case we need to pick 14 of the 16 coin tosses to be the ones that ended up heads. There are  $C(16, 14) = (16 \times 15)/2 = 120$  different sequences that have 14 heads and 2 tails. So in this case,  $P(X = 14 \text{ heads}) = \sum_{\text{all sequences}} P(H)^{14} \times P(T)^2$  or

$120 \times \pi^{14} (1 - \pi)^2$  where  $\pi$  is the probability of heads. When we assume that the coin is equally likely to land on heads or tails ( $\pi = 0.5$ ), we find that the probability of getting 14 heads is  $120 \times 0.5^{16} = 0.00183$ .

In general, the probability of obtaining  $k$  successes in a sequence of  $n$  trials with success probability  $\pi$  on each trial is:  $P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$  where  $\binom{n}{k} = \frac{n!}{k!(n - k)!}$

It can also be shown that expected value of a binomial random variable is  $E(X) = n\pi$  and the variance is  $V(X) = n\pi(1 - \pi)$  when  $X$  is  $\text{Binomial}(n, \pi)$ . (See Exploration B for more discussion of expected value and variance of a random variable.)

## An Alternative Measure of Rareness

As an alternative to the p-value, another measure of where an observation falls in a distribution is “how many standard deviations” it is from the mean of the distribution. But if our data is “helper” or “hinderer,” how do we calculate the standard deviation? Recall that the standard deviation measures the variability or spread of a distribution from the mean and refers to how clustered together or *consistent* the values are.

Suppose we had “coded” each infant’s outcome as 0 or 1:

ID	Infant choice	Coded choice									
1	Helper	1	5	Helper	1	9	Helper	1	13	Helper	1
2	Helper	1	6	Helper	1	10	Helper	1	14	Helper	1
3	Hinderer	0	7	Hinderer	0	11	Helper	1	15	Helper	1
4	Helper	1	8	Helper	1	12	Helper	1	16	Helper	1

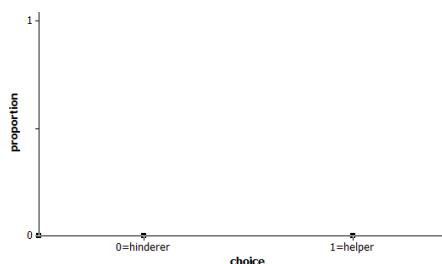
(e) If we treat the data numerically in this manner, what is the mean or average of these 16 values?

(f) Use technology (e.g., [Descriptive Statistics](#) applet) to calculate the standard deviation of these 16 values.

**Discussion:** We are again looking at the “typical deviation” from the mean. With 14 ones and two zeros, the mean is  $14/16 = 0.875$  (the same as the sample proportion). The squared deviations from this mean are  $(1 - 0.875)^2$  and there are 14 of these (so most of the data values are pretty close to the mean) and  $(0 - 0.875)^2$  and there are two of these. So summing these squared deviations and dividing by  $n - 1 = 15$ , we get a variance of  $1.75/15 \approx 0.1167$ . Taking the square root of the variance gives us the standard deviation.

(g) Take the square root of this variance, how does it compare to the standard deviation calculated in (f)?

(h) Suppose 8 infants had chosen the helper toy and 8 infants had chosen the hinderer toy, sketch the “dotplot” in this case.



(i) Would you consider those results more, or less, or equally “consistent” than the actual results?

(j) Use technology to calculate the standard deviation (by hand or using technology) for this hypothetical 8/8 split. Does this confirm your answer to (i)? In other words, which has more variability, the hypothetical data or the actual study? (In other words, in which case would it be easier to predict a future observation?)

(k) Is it possible for a data set of 16 zeros and ones to have a *larger* standard deviation than in (j)? What would need to be true for the 16 values to have the *smallest* possible standard deviation (and what is the value of the standard deviation in that case)?

Largest:

Smallest:

**Discussion:** If data tend to be far from the average value, then we consider these data to be “less consistent” or “more spread out” and we say the data set has “more variability.” It’s critical to keep in mind that variability refers to (horizontal) distances of the data values from the center of the distribution. If most values are close to the mean, the variability will be small, and so the average deviance from the center will be small. If many of the values are far from the mean of the data set, then the variability will be larger, and so the average deviation will be larger. Although it’s actually rather unusual to consider variability when we have only yes/no outcomes (it really doesn’t give us any additional information beyond the sample proportion), we hope this discussion will help you keep in mind this interpretation of the concept of variability and how average deviation from the mean provides a reasonable measure of variability.

(l) But in this case, we actually want to know about the variance of our random variable, *number of babies that choose the helper toy*. The Probability Detour for the Binomial distribution tells us that  $V(X) = n\pi(1 - \pi)$  for different repetitions of the random binomial process. For what values of  $\pi$  will this variance be the largest? Smallest? (*Hint:* You may want to graph this function for different values of  $\pi$ ).

(m) Calculate the standard deviation of this probability model under the null model (assuming  $\pi = 0.50$ ).

(n) Now calculate the number of standard deviations 14 falls from the center of the null distribution.

We still need a way to judge whether your answer to (n) is considered large. We will discuss this in more detail later, but for now, when we have a fairly symmetric distribution, as your null distribution should be in this investigation, values larger than 2 or less than -2 are typically considered extreme.

(o) Is your answer to (n) larger than 2? Is this consistent with your p-value for this study? Explain.

### Practice Problem 1.1B

- (a) Suppose we thought there was a 60% chance that an infant would choose the helper toy. Explain how we would change our simulation to represent this null model.
- (b) Carry out a simulation (e.g., One Proportion Inference applet) to determine whether our observed result of 14 helper choices by 16 infants is surprising for a process with  $\pi = 0.60$ . Report and evaluate an approximate p-value.
- (c) How does the distribution of the “number of successes” with  $\pi = 0.60$  compare to when  $\pi = 0.50$ ? Comment on shape, center, and variability.
- (d) Determine the exact p-value using the binomial distribution. How does it compare to your simulation results?
- (e) For the binomial model, report the expected value and standard deviation of the number of infants choosing the helper toy with  $\pi = 0.60$  and  $n = 16$ .
- (f) How many standard deviations is 14 from the expected value for this model? Is it larger or smaller than what you found in (n) in the previous exploration? Explain why this change makes sense.

### Technology Detour – Loading in a Data File

- In RStudio, choose Import Dataset > From Web URL and enter the URL, then press Import.
  - e.g., <http://www.rossmanchance.com/iscam3/data/InfantData.txt>
  - Keep Heading selected as Yes etc.
  - Note: With other data files, you may need to consider how “missing values” are coded.

#### • In R

*Files from the web:* To load this data into R, open the InfantData.txt (raw data) link from the data files page, select all, copy, and then in R use the following command (Keep in mind that R is case sensitive):

**PC:** > InfantData = read.table("clipboard", header=TRUE)

**MAC:** > InfantData = read.table(pipe("pbcpaste"), header=TRUE)

You can also use a URL (in quotes). The header command indicates the variables have names.

*Txt files on your computer:*

```
> InfantData = read.table(file.choose(), header=T)
```

To see the data, type > InfantData or > View(InfantData) or > head(InfantData)

Next you can “attach” the file to be able to use variable names directly:

> attach(births) - Now R knows what the “choice” variable is

If you don’t “attach” the file then you need to clarify to R which datafile you are using (e.g., InfantData\$choice).

Other input options, depending on how the data you are pasting is formatted, include:

sep="\t" - separated by tabs

na.strings="\*" - how to code missing values

dec=". " - decimals

## Technology Detour – Tallying the Outcomes

**In R:** To count the number of correct and incorrect responses, type:

> summary(InfantData) or > table(InfantData)

## Technology Detour – Bar Graphs

### Creating a Bar Graph in R (raw data)

- You need to pass the tabled data into the barplot function:  

```
> barplot(table(InfantData), xlab="Choice", ylab="Frequency")
```
- You may need to toggle to the R Graphics Window to see the graph window. Now you can use R to export the graph to a file or you can Copy and Paste or use a screen capture.

### Creating a Bar Graph in R (summary data)

- If you already (or only) have the summarized data (the number of successes and failures), can type:  

```
> barplot(c(14, 2), names.arg=c("Helper", "Hinderer"))
```

## Technology Detour – Calculating Binomial Probabilities

### Calculating Binomial Probabilities in R

- Make sure you have the ISCAM Workspace loaded.
- The iscabinomprob function takes the following inputs
  - *k* = the observed value you want to calculate the probability about
  - *n* = the sample size of the binomial distribution
  - *prob* = the probability of success in the binomial distribution
  - *lower.tail* = TRUE if you want the probability less than or equal to the observed value, FALSE if want the probability greater than or equal to the observed value
- For the Friend or Foe study, in the Console at the prompt, type  
`> iscabinomprob(k=14, n=16, prob=.5, lower.tail=FALSE)`  
 or `> iscabinomprob(14, 16, .5, FALSE)`  
 (Yes, “false” needs to be capitalized)

You should see both the probability appear in the Console window and a shaded graph open in a separate Graphics window. (You may have to toggle windows to see this Graphics window.)

## Handy Reminders

- **In R:** If you don’t remember the inputs for an ISCAM R Workspace function, just pass “?” (with quotations marks) into the function, e.g., `> iscabinomtest("?)`. When the text lists terms for the inputs (in italics), those can be used to identify the inputs and then you may enter them in any order. Also, if you want to return to earlier commands, you can use the up arrow to scroll through them.

### Investigation 1.2: Do Names Match Faces?

A study in *Psychonomic Bulletin and Review* (Lea, Thomas, Lamkin, & Bell, 2007) presented evidence that “people use facial prototypes when they encounter different names.” Similar to one of the experiments they conducted, you will be asked to match photos of two faces to the names Tim and Bob. The researchers wrote that their participants “overwhelmingly agreed” on which face belonged to Tim. You will conduct a similar study in class to see whether your class also agrees with which face is Tim’s more than you would expect from random chance (if there is no facial prototyping and people pick a name for the face on the left at random). (Note, you may want to alternate or randomize which face is placed on the left, which name you mention first, etc.)

(a) Describe in words the null model for this study.

(b) Consider whether we can model this study as a binomial random process. Do you consider the properties necessary for a binomial random variable to be met in this context? What assumptions need to be made? [Hint: See the Probability Detour box in the previous investigation.]

### Summarizing the Data

Your instructor will collect and share your class results.

(c) How many subjects participated in your study? What proportion of the respondents agreed that the face with the goatee was Tim? Does this appear to be an “overwhelming majority” to you?

(d) Produce a bar graph of the results. Make sure the axes are well-labeled. Can we conclude from this graph that people use facial prototyping? Explain.

## Terminology Detour

To improve our communication about data, we need to define several terms that will be used throughout this book.

**Definition:** The observational units are the people or objects about which data are recorded. A variable is any characteristic that varies from observational unit to observational unit. A study's sample size is the number of observational units in the study.

In Investigation 1.1, the observational units were the 16 infants. The variable of interest was which toy the child chose to play with. Other variables we could have recorded on the infants include age, handedness, and gender.

**Definition:** We classify the type of variable as categorical (assigning each observational unit to a category) or quantitative (assigning each observational unit a numerical measurement). A special type of categorical variable is a binary variable, which has just two possible outcomes.

One way to decide between these types is to ask yourself whether or not you can take the average of the values. For example, gender is categorical but age is quantitative. The zip code of the baby's home would be categorical even though it is a number, because arithmetic operations like adding or averaging do not make sense on zip codes. Age can often be treated either way. We will consider it quantitative if the data were recorded in such a way that it makes sense to discuss the average age of the employees, but not if they have been grouped into age categories.

Also keep in mind that when you have a binary variable, we often (and it may be arbitrarily) define one outcome to be a "success" (the one we will focus on) and the other to be a "failure."

**Definition:** We will also differentiate between a statistic and a parameter. A *statistic* is a (known) numerical value that summarizes the observed results, whereas a *parameter* is the same numerical summary but applied to the underlying random process that generated the data (whose value is usually unknown).

In Investigation 1.1, the statistic could be either the number (14) or the proportion (0.875) of infants who chose the helper toy. The parameter would then be the long-run probability of an infant picking the helper toy. (We rejected 0.5 as the value of this parameter, deciding it's actually larger than 0.50.)

With a single categorical variable, we will refer to a sample proportion (a statistic) by the symbol  $\hat{p}$  and a *process probability* (a parameter) by the symbol  $\pi$ .

(e) Define the observational units and variable for this study. Is the variable quantitative or categorical? If categorical, is it binary? If binary, define a success and a failure.

(f) Define a statistic and a parameter for this study (and denote each by the appropriate symbols). For which of these (parameter or statistic) do we know the value, and for which do we not know the value? For the one that we know, report the value as well.

Statistic:

Parameter:

## Making Conclusions Beyond the Data

(g) If there is nothing to the theory that people associate names with faces, what does this imply about our parameter, the probability of picking Tim for the name of the face with the goatee? That is, what is the parameter *value* under the null model?

If we use the symbol  $\pi$  to represent the probability of success, we can use short-hand notation to refer to this null model or [null hypothesis](#):

Null Hypothesis:  $\pi = 0.5$

(h) If the research conjecture is correct that people are more likely to associate the name Tim with the face with the goatee, what does that imply about the probability of success?

We can translate this research conjecture into the [alternative hypothesis](#) or, in shorthand,

Alternative hypothesis:  $\pi > 0.5$

Typically, the alternative hypothesis is a statement of the research conjecture, what the researchers hope to establish from the study. In contrast, the null hypothesis is the uninteresting version of the conjecture, saying there really is nothing going on, the “dull” hypothesis.

Defining the alternative hypothesis is important because this determines which direction we will look in when we calculate the p-value. That is, the direction of the alternative hypothesis (e.g., greater than or less than) determines which results we will consider “or more extreme” than the observed result.

**Notice** that hypotheses are always about *parameters*, not statistics. We do not need to test a hypothesis about the value of a statistic, because we calculate that value from the sample data. Also notice that the hypotheses can (and should) be stated *before* collecting or analyzing the data.

(i) Describe how we could simulate “could have been” observations for this study under the assumption of the null hypothesis. (*Hints*; Can you use a coin? Why? How many tosses?)

(j) Describe how you would use the binomial distribution to calculate the exact p-value for this study. Be sure to specify the values of  $n$  and  $\pi$ , and also the values that you would calculate the probability of.

This process of evaluating the strength of evidence against the null hypothesis is generally referred to as a *test of significance*. When the p-value is computed from the Binomial distribution, this is often called a *Binomial Test*. Many statistical software packages will carry out a Binomial test directly.

## Technology Detour – Binomial Test of Significance

### Conducting a Binomial Test in R (Summarized Data)

The `iscbinomtest` function takes the following inputs:

- *observed* = Observed number of successes or proportion of successes
  - With a data vector, can first determine number of successes, e.g., `table(NamesData)`
  - If you enter a value less than one, it will assume you entered the proportion
- *n* = Number of trials (sample size)
- *hypothesized* = Hypothesized probability
- *alternative* = Direction of alternative (e.g., “greater” or “less”)

For the Friend or Foe study, using the command (with or without input labels)

```
iscbinomtest(observed=14, n=16, hyp=.5, alt="greater")
```

should show output in the Console window as well as a graph of the binomial distribution with the p-value shaded in the Graphics window.

(k) Find and report the p-value from your technology (including appropriate notation for the event you used).

(l) Write a one-sentence *interpretation* of the p-value that you have calculated. Be sure it is clear what you found the probability of and any assumptions underlying the calculation. [Hint: Recall the simulations you performed in the Investigation 1.1.]

(m) Would you consider the class results to be *statistically significant*? Explain.

## Conclusions

(n) Based on your calculations, does your class study provide strong evidence that students are likely to match the name Tim to the face with the goatee more than we would expect if they are just guessing?

### Study Conclusions

With a small p-value, we would have very strong evidence to *reject the null hypothesis* and conclude that the students like those in your class have a genuine tendency to identify the “correct” name with the picture of Tim. Again, we might be cautious in generalizing these results to other names/faces or even to other subjects identifying the names. We have also assumed that students behave as a binomial process and all have the same probability of picking Tim (the randomness is in each person’s selection) and that they don’t influence each other’s choices. If you did alternate or randomize the order of the faces or names, then you would feel comfortable thinking that which one ended up on the left was not the reason most students picked Tim to match the picture with a goatee.

### Practice Problem 1.2

In 2011, an article published by the medical journal *Gut – An International Journal of Gastroenterology and Hepatology* (Sonoda, et al.) reported the results of a study conducted in Japan in which a dog was tested to see whether she could detect colorectal cancer. The dog used was an 8-year-old black Labrador named Marine. (As her name might suggest, she was originally trained for water rescues.) The study was designed so that the dog first smelled from a bag that had been breathed into by a patient with colorectal cancer. This was the standard that the dog would use to judge the other bags. Marine then smelled the breath in five different bags from five different patients, only one of which contained breath from a colorectal cancer patient (not the same as the original patient); the others contained breath from non-cancer patients. The dog was then trained to sit next to the bag which she thought contained breath from a cancer patient (i.e., had the cancer scent). If she sat down next to the correct bag, she was rewarded with a tennis ball. Marine completed 33 attempts of this experimental procedure, with a different set of five patients each time: four non-cancer patients and one cancer patient. And, each time, whether or not she correctly identified the bag with the breath of the cancer patient was recorded.

- (a) Identify the observational units [*Hint: What is the sample size?*] and variable in this study. Is the variable quantitative or categorical?
- (b) Is it reasonable to model this study as a binomial process? Justify your response for each condition.
- (c) Write a one-sentence interpretation of the *parameter* of interest in this study. What symbol can we use to refer to this parameter?
- (d) Write a one-sentence interpretation of the *statistic* we will observe in this study. What symbol can we use to refer to this statistic?

### Investigation 1.3: Heart Transplant Mortality

Poloneicki, Sismanidis, Bland, and Jones (2004) reported that in September 2000 heart transplantation at St. George's Hospital in London was suspended because of concern that more patients were dying than previously. Newspapers reported that the 80% mortality rate in the last 10 cases was of particular concern because it was over five times the national average. The variable measured was whether or not the patient died within 30 days of the transplant. Although there was not an officially reported national mortality rate (probability of death within 30 days for patients undergoing this procedure), the researchers determined that 15% was a reasonable benchmark for comparison.

- (a) Define the observational units and variable for this study. Is the variable quantitative or categorical?

Observational units:

Variable:

Type:

With a binary categorical variable, we need to define which outcome we will consider “success” and which we will consider “failure.” The choice is often arbitrary, though sometimes we may want to focus on the more unusual outcome as success. In fact in many epidemiology studies, “death” is typically the outcome of interest or “success.”

- (b) Considering death within 30 days as a success, define the parameter of interest in this study (in words), including the symbol we can use to refer to it. Is it reasonable to model this heart transplantation process as a binomial process?

Parameter:

Binomial process?

- (c) If there is nothing unusual about the mortality rate for transplants at this hospital (compared to other U.K. hospitals), what does this imply about the probability of “success”?

- (d) If the patients at this hospital are indeed dying at a higher rate than the benchmark rate, what does this imply about the value of the probability of “success”?

- (e) Translate your answers to (c) and (d) to a null and alternative hypothesis statements. Keep in mind, the null hypothesis claims the observed result is “just by chance,” whereas the alternative hypothesis translates the research conjecture.

Null hypothesis (often denoted  $H_0$ ):

Alternative hypothesis (often denoted  $H_a$ ):

(f) Of the hospital's ten most recent transplantations at the time of the study, there had been eight deaths within the first 30 days following surgery. Is this sample result in the direction suspected by the researchers? Explain. What symbol could we use to refer to this proportion?

(g) Describe two methods for obtaining a p-value to assess whether this observed result is statistically significant. (Be sure it's clear what is different in this scenario from what you did in previous investigations.)

**Recap:** We have seen two main ways to obtain a p-value to assess the strength of evidence these data provide against the claim that the mortality rate at this hospital could be 0.15. What's different now is that we are interested in testing a null probability other than 0.5.

#### Approach 1: Simulation

Assume the mortality rate is 0.15 and generate thousands of samples of 10 patients, counting the number that die in each sample. Examine this null distribution to see how unusual it is to find 8 or more patients dying in the sample by chance alone. (Previously, we looked at both physical coin tossing and the One Proportion Inference applet as ways to simulate these outcomes with a success probability of  $\pi = 0.5$ . Now we can't use coins but we could use spinners for a physical model of this random process.)

#### Approach 2: Exact binomial probabilities

Use binomial probabilities to calculate the long-run relative frequency of 8 or more deaths from a binomial process with  $n = 10$  and  $\pi = 0.15$ . That is, determine  $P(X \geq 8)$  where  $X$  represents the number of successes. You could calculate this probability by hand using the formula or with technology (R (either `iscabinomprob` or `iscabinomtest`), or Minitab, or the One Proportion applet).

With either approach, a small p-value provides evidence against the null hypothesis underlying the calculation (we "reject the null hypothesis" as being plausible based on what we observed). If you examine enough samples in the simulation, your result will be quite close to the exact binomial approach.

- (h) Open the [One Proportion Inference](#) applet and generate 1000 samples from a binomial process with  $\pi = 0.15$  (under the null model) and  $n = 10$ . Comment on the shape of the distribution.

Probability of success ( $\pi$ ):	0.15
Sample size( $n$ ):	10
Number of samples:	1000

- (i) Estimate the p-value from this null distribution. Clearly explain how you did so.

(j) Check the **Exact Binomial** box to have the applet determine the exact p-value. Report and interpret this value below (what is it the probability of?).

(k) Based on this analysis, what conclusions would you draw about the probability of death within 30 days of a heart transplant at this hospital?

### Does it matter which outcome I choose to be success?

(l) Suppose that we had focused on survival for 30 days rather than death within 30 days as a “success” in this study. Describe how the hypotheses would change and how the calculation of the binomial p-value would change. Then go ahead and calculate the exact binomial p-value with this set-up. How does its value compare to your answer in (j)? How (if at all) does your conclusion change?

### Does the sample size matter?

(m) Following up on the suspicion that the sample of size 10 aroused, these researchers proceeded to gather data on the previous 361 patients who received a heart transplant at this hospital dating back to 1986. They found 71 deaths.

Calculate the sample proportion for these data:

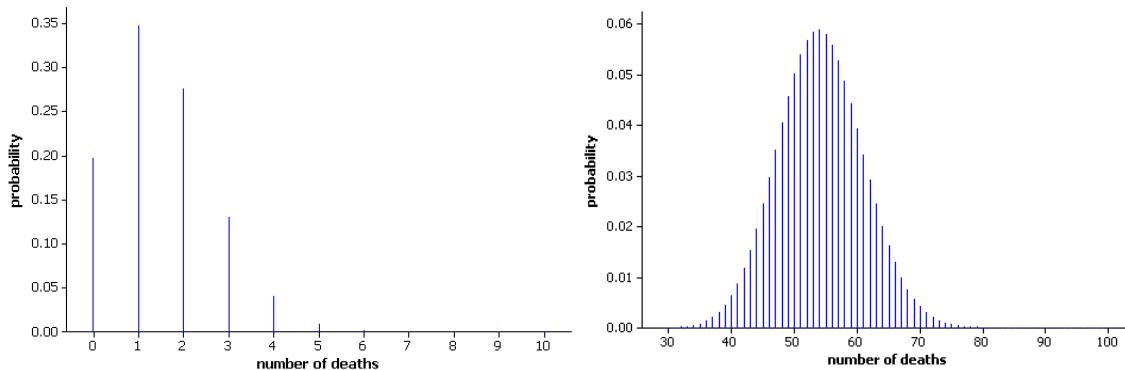
Predict whether this is more or less convincing evidence that this hospital’s death rate exceeds 0.15? Explain your reasoning.

Use the [One Proportion Inference](#) applet to determine (approximately with simulation, and exactly with the binomial distribution) the probability of finding at least 71 deaths in a sample of 361 if  $\pi = 0.15$ .

(n) Is the probability in (m) convincing evidence to consider the sample result surprising if the mortality rate at this hospital matched the national rate? Explain.

(o) Is the evidence against the null hypothesis stronger or weaker than the earlier analysis based on 10 deaths? Explain how you are deciding and why the strength of evidence has changed in this manner.

The following graphs display the two theoretical probability distributions (for sample sizes  $n = 10$  and  $n = 361$ ), both assuming the null hypothesis ( $\pi = 0.15$ ) is true. These graphs show just how far the observed values (8 and 71) are from the [expected value](#) of the number of deaths ( $0.15 \times 10 = 1.5$  and  $0.15 \times 361 = 54.15$ ) in each case. You should also note that the shape, center, and variability of the probability distribution for number of successes are all affected by the sample size  $n$ .



Keep in mind that of interest to us is our observation's relative location in the null distribution. Thus, we are most interested in how variable the possible outcomes are from the “expected” outcome. The center of the distribution isn't all that interesting to us in answering the research question because we determine what the center of the distribution will be by how we specify the null model. Even the shape isn't all that interesting for its own sake in answering our research question.

(p) Identify one other feature of the above distributions that differs between them.

- (q) Which data set do you think is more valid to use – the larger sample size or the more recent data? Explain how you are deciding.

### Study Conclusions

A sample mortality rate of 80% is indeed quite surprising, even with a sample size as small as 10, if the actual probability of death were 0.15. The (exact) p-value is 0.0000087, providing very strong evidence that the actual probability of death at this hospital is higher than the national benchmark of 0.15. However, we must be cautious about doing this type of “data snooping,” where we allowed a seemingly unusual observation to motivate our suspicion and then use the same data to support our suspicion. Once the initial suspicion has formed, we should collect new data on which to test the suspicion. The above investigation examined all previous heart transplantations at this hospital over the previous 14 years. In this broader study, the p-value is 0.0097, still providing strong evidence against the null hypothesis. That is, there is strong evidence that this hospital’s probability of mortality was higher than the 15% benchmark. We must, however, be cautious because our study has not identified what factors could be leading to the higher rate. Perhaps this hospital tends to see sicker patients to begin with. The researchers actually performed a more sophisticated analysis that incorporated information about the risk factors of all the operations at this hospital and reached similar conclusions.

### Practice Problem 1.3

Reconsider Practice Problem 1.2. Marine was correct in 30 of the 33 attempts.

- (a) State the null and alternative hypotheses for testing whether Marine is more likely to pick the correct breath sample than if she was just randomly guessing between the five bags each time.
- (b) Determine the p-value for Marine and provide a detailed *interpretation* of the p-value you find.
- (c) Summarize the conclusions you would draw from this study: Do you think Marine got lucky or do you think something other than random chance was at play? How strong is the evidence?
- (d) Suppose in another study, a different dog was correct 60 times. Would you consider that convincing evidence? Explain.

### Investigation 1.4: Kissing the Right Way

Most people are right-handed and even the right eye is dominant for most people. Researchers have long believed that late-stage human embryos tend to turn their heads to the right. German biopsychologist Onur Güntürkün (*Nature*, 2003) conjectured that this tendency to turn to the right manifests itself in other ways as well, so he studied kissing couples to see if both people tended to lean to their right more often than to their left (and if so, how strong the tendency is). He and his researchers observed couples from age 13 to 70 in public places such as airports, train stations, beaches, and parks in the United States, Germany, and Turkey. They were careful not to include couples who were holding objects such as luggage that might have affected which direction they turned. We will assume these couples are representative of the overall decision making process when kissing.

- (a) Identify the observational units, the variable, and statistic in this study. Is the variable quantitative or categorical? Also define the statistic and the parameter in words. What symbols do we use to refer to the statistic and parameter?

Observational units:

Variable:

Type:

Statistic:

Symbol:

Parameter:

Symbol:

- (b) Dr. Güntürkün noted that a strong majority of people have a dominant right foot, or eye, and conjectured that people would exhibit a similar tendency of “right sidedness” when kissing. Our parameter of interest, denoted by  $\pi$ , is the probability that a kissing couple that is selected at random (from a public place in the United States, Germany, and Turkey) leans to the right. A recent (2012) Tic Tac candy ad claimed that 74% of people tilt their head to the right when they kiss. Suppose that before seeing the data Dr. Güntürkün conjectured the probability of leaning right to be equal to 0.74.

- (c) Do you think 0.74 is too large, or too small, or correct?

- (d) Translate the Tic Tac claim into a null hypothesis statement using appropriate symbols,

$H_0$ :

- (e) Although we may suspect most couples to lean to the right, we don’t have an obvious conjecture for whether the probability is larger or smaller than 0.74. In this case, we can test whether the process probability *differs* from 0.74. Translate this into an alternative hypothesis statement.

$H_a$ :

With such a two-sided alternative hypothesis (rather than a “one-sided” alternative of strictly less than or strictly greater than), we calculate the p-value as the probability of obtaining a result as extreme as we did *in either direction* (in both the right and left tails of the null distribution).

- (f) Suggest a method for calculating such a p-value.

**Definition:** [Two-sided p-values](#) are used with two-sided alternative hypotheses (not equal to). A two-sided p-value considers outcomes in both tails that are at least as rare as the observed result, where at least as rare could imply a smaller probability of occurring or being as far or farther from the expected value. If the null distribution is symmetric these two approaches are equivalent and the two-sided p-value will be double the one-sided p-value.

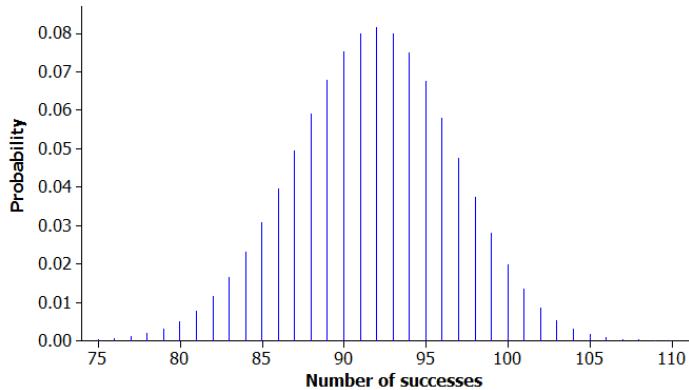
In total, 124 kissing pairs were observed with 80 couples leaning right (*Nature*, 2003).

(g) To determine which values are more extreme than 80:

How many of the 124 couples would you expect to turn to the right if the process probability equals 0.74?

Is 80 more or less than this expected value?

What would be a reasonable value to use on the *other* side of the null distribution, for providing evidence against this null hypothesis? Explain.



(h) Use the [One Proportion Inference](#) applet to estimate the two-sided p-value:

- Specify the appropriate values of  $\pi$  and  $n$  for this study.

$\pi =$                            $n =$

Probability of success ( $\pi$ ):

Sample size ( $n$ ):

Number of samples:

- Enter a large number for **Number of samples**
- Press **Draw Samples**
- For **Number of successes**, enter 80 but toggle the button to change the  $\geq$  to  $\leq$
- Check **Two-sided**
- Check **Exact Binomial**

Animate

**Draw Samples**

Number of successes

Proportion of successes

As extreme as   Count

---

Two-Sided

Exact Binomial

The applet found a p-value by adding two tail probabilities together; what were the cut-off values for these tails? How do the cut-offs compare to what you anticipated above? What value does it report for the exact two-sided p-value?

- (i) What conclusion do you draw from this p-value about whether the probability of a kissing couple leaning to the right is 0.74? Explain briefly.

Strong evidence against the null hypothesis?

Why are you making this decision?

- (j) Dr. Güntürkün actually conjectured  $2/3$  as the process probability of a kissing couple leaning to the right (consistent with some other right-sided tendencies by humans). Repeat (h) to determine a two-sided p-value for testing this hypothesis. Report the p-value and summarize your conclusion about the plausibility that  $\pi = 2/3$ .

[Hint: In R, use `iscambinomtest`, with `alternative="two.sided"` (in quotes, with the period).]

- (k) Repeat your analysis method from (j) to determine whether 0.60 is a plausible value for the probability that a kissing couple leans to the right.

### Study Conclusions

These sample data provide strong evidence against the null hypothesis that kissing couples lean to the right 74% of the time (two-sided p-value = 0.0185), but we are not able to reject the null hypothesis that the probability of turning to the right in this kissing process is equal to  $2/3$  (two-sided p-value = 0.6341) or 0.60 (p-value = 0.3151). It is plausible that this sample result (80 successes out of 124 trials) came from a process where the probability of success for the process was  $2/3$ . But it is also plausible that the sample result came from a process where the probability of success was 0.60. Note that we haven't *proven* either of these values to be the correct one. The sample data allow for many (technically, infinitely many) plausible values of this probability, as you will explore in the next investigation. We should be cautious in generalizing these results to all kissing couples. Based on the study description, the researcher did attempt to get a broad representation of couples, but there could be some hidden sources of bias.

- (l) Return to the case where  $\pi = 0.74$ . We found the cut-offs for the two-sided p-value to be 80 (our observed result) and 104 (as being just as extreme on the other side of the expected value). Use 104 as the observed count in the applet. What value does the applet use for the lower cut-off, still 80?

## Technical Details

In calculating a two-sided p-value in the previous investigation, we considered the values that were as extreme as 80 “or more extreme” which we interpreted to be “as far from the expected value.” When  $\pi = 2/3$ , the expected number of successes out of 124 is 82.67 so that the observed value of 80 successes in the sample is 2.67 below this expected value. We can also find  $P(X \leq 80)$  to be 0.336 using the binomial distribution. To find the tail probability in the other direction, you can first determine the values that are as extreme in the other direction,  $82.67 + 2.67 = 85.33$ , and then you would calculate  $P(X \geq 85.33) = P(X \geq 86) = 0.298$ . So then the two-sided p-value reported by the applet was  $P(X \leq 80) + P(X \geq 86) = 0.336 + 0.298 = 0.634$ .

Another way to interpret “or more extreme” is to consider the outcomes that are even less likely to occur than the value observed. That is, we will include an observed value of  $x$  in our p-value calculation if the probability for that value is smaller than the probability of our observed value. In some situations, this will lead to slightly different results. Below is a portion of the binomial distribution with  $\pi = 2/3$ :

$x$	79	80	81	82	83	84	85	86	87
$P(X = x)$	0.0581	0.0655	0.0712	0.0748	0.0759	0.0742	0.0699	0.0635	0.0556

Here,  $P(X = 80) = 0.0655$ ,  $P(X = 85) = 0.0699$ ,  $P(X = 86) = 0.0635$  so that 86 is the first value of  $x$  that has a smaller probability than 80. This alternative method then tells us to report the two-sided p-value as  $P(X \leq 80) + P(X \geq 86)$ . However, if we had started with 86, then the two-sided p-value will be calculated from  $P(X \geq 86) + P(X \leq 79)$  as 79 is the first  $x$  value with a probability smaller than  $P(X = 86)$ . When the sample size is large and the probability of success is near 0.5, these two methods should lead to very similar results. However, this latter approach (called the *method of small p-values*) accounts for the non-symmetric shape that the binomial distribution often has with small  $n$  and  $\pi$  away from 0.5. This is the method used by R. Though we will not advocate one approach over the other here, you will often want to be aware of the algorithm used by your statistical package. In fact, Minitab uses a third approach altogether. Some other software packages find the (smaller) one-sided p-value and simply double it. We find this approach less satisfying when the binomial distribution is not symmetric.

## Practice Problem 1.4

- (a) Do these data provide convincing evidence that the probability of leaning right is larger than 0.5? (State hypotheses, report the p-value, describing how you determined it and clarifying what it is in terms of the probability of  $P(X \dots)$ , and interpret the strength of evidence against the null hypothesis.)
- (b) Do these data provide convincing evidence that the probability of leaning right differs from 0.5?
- (c) What two values are used in (b) for the cut-offs and how far is each from the expected value of  $X$ ?

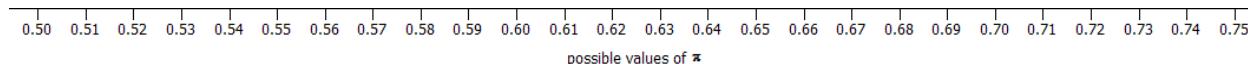
### Investigation 1.5: Kissing the Right Way (cont.)

In the previous investigation, you learned how to decide whether a hypothesized value of the parameter is plausible based on a two-sided p-value. The two-sided p-value is used when you do not have a prior suspicion or interest in whether the hypothesized value is too large or too small. In fact, in many studies we may not even really have a hypothesized value, but are more interested in using the sample data to estimate the value of the parameter.

- (a) Considering that the researchers observed  $\hat{p} = 80/124$  couples lean to the right, what is your best (single value) guess for the value of  $\pi$ , the underlying probability that a kissing couple leans to the right?
- (b) Do you believe the value of the parameter  $\pi$  is exactly equal to the value specified in (a)? Do you think it is close? How close?

We can employ a “trial-and-error” type of approach to determine which values of  $\pi$  appear plausible based on what we observed in the sample. This involves testing different values of  $\pi$  and seeing whether the corresponding two-sided p-value is larger than some pre-specified cut-off, typically 0.05. (This cut-off is often called the [level of significance](#).) That is, we will consider a value plausible for  $\pi$  if it does not make our sample statistic look surprising.

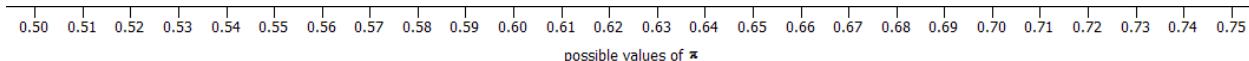
- (c) You found in the previous investigation that 0.74 does not appear to be a plausible value for  $\pi$ , but 0.6667 does because the two-sided p-value for testing  $\pi = 0.6667$  is larger than 0.05. Use the [One Proportion Inference](#) applet to determine the values of  $\pi$  such that observing 80 of 124 successes or a result more extreme occurs in at least 5% of samples. [Hints: Use values of  $\pi$  that are multiples of 0.01 until you can find the boundaries where the exact two-sided p-values change from below 0.05 to above 0.05. Then feel free to “zoom in” to three decimal places of accuracy if you’d like.] On the number line below, indicate with an X which values are rejected and with an O which values are not rejected and therefore considered plausible.



**Definition:** A [confidence interval \(CI\)](#) specifies the plausible values of the parameter based on the sample result.

Confidence intervals are an additional or alternative step to a test of significance (p-value) that tells us whether or not we have strong evidence against one particular value for the parameter. Again, statistical software packages employ slightly different algorithms for finding the binomial confidence interval. This investigation should help you understand how to interpret the resulting confidence interval given by technology. What you found in (c) will be called a “95% confidence interval” as it was derived using the  $1 - 0.95 = 0.05$  cut-off value.

- (d) Repeat (c) but using 0.01 rather than 0.05 as the criterion for rejection/plausibility. [Hints: You can check the **Show sliders** box and use the slider or edit the orange number to change the value of  $\pi$ .] Does this “99% confidence interval” include more or fewer values than the one based on the 0.05 criterion? Explain why this makes intuitive sense.



Explanation:

### Technology Detour – Binomial Confidence Intervals

**In R:** Specify a confidence level in the function:

```
> iscambinomtest(observed = 80, n = 124, conf.level=.95)
```

[You do not have to specify a hypothesized value or alternative direction with this function, but you do then need to label the confidence level input because R will otherwise assume you are telling it a hypothesized value, the next input in the list. This format will also produce a graphical display of the confidence interval.]

- (e) Use technology to verify the confidence interval endpoints that you found in (c).

- (f) Use technology to determine a 99% confidence interval for the probability that a kissing couple leans to the right. Comment on how this interval compares to the 95% interval, examining both midpoints and widths. Also indicate whether your answer to (d) turned out to be correct.

- (g) Interpret the confidence interval from (f). What are you 99% confident of?

## Study Conclusions

The researchers are assuming they have a representative sample from a process and want to determine  $\pi$ , the underlying probability that a randomly selected kissing couple leans to the right. Based on this sample of 124 observations, we estimate  $\pi$  to be close to  $80/124 = 0.645$ . However, we know there is some sampling variability, so we want to find an interval of values that appear to be plausible values of  $\pi$ . We do this by finding the values of  $\pi$  for which the two-sided p-value is greater than 0.05. These are all the values of the parameter such that our sample result is not overly surprising. You should have found this “95% confidence” interval to be approximately 0.554 to 0.729 (results from using Minitab or R or the applet will differ slightly). Thus, based on these sample results, we are “confident” that the actual value of  $\pi$ , the probability a random kissing couple leans right, is between 0.55 and 0.73. A 99% confidence interval for  $\pi$  extends from 0.526 and 0.752 and therefore includes more values than a 95% interval. The higher level of confidence requires more “room for error.” You will learn more about confidence intervals in the next section.

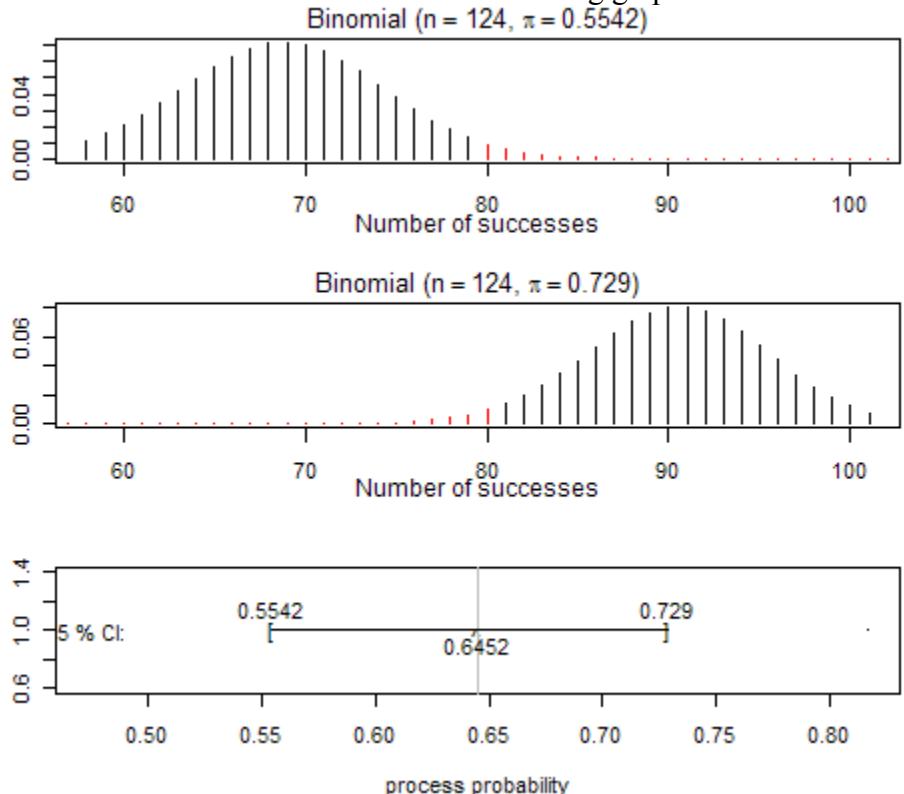
**Discussion:** In this investigation you have learned a second type of “statistical inference”: based on the sample statistic, providing an interval of plausible values for the parameter. Confidence intervals provide a nice companion to tests of significance and are also very useful by themselves. Whereas a test of significance allows to you test a specific hypothesized value, if you reject the null hypothesis, the test of significance provides no information as to *how different* the actual parameter is from the hypothesized value. The confidence interval provides an estimate (with bounds) of the actual value of the parameter.

In fact, there is a type of [duality](#) between confidence intervals and tests of significance. The confidence interval is the set of values for which we would *fail to reject* the null hypothesis in favor of the *two-sided* alternative. In fact, Minitab’s algorithm for the two-sided p-value is obtained by maintaining this correspondence (the small p-values approach may lead to small departures from this duality). So we can interpret the confidence interval as the set of plausible values for the parameter in that they are the values such that our observed sample result would not be surprising.

In summarizing your results, remember to always conclude with an answer to the research question *in context*. The decision to reject or fail to reject  $H_0$  should never be your last statement. Similarly, the determination of a confidence interval should never be the last statement. Include at least one more sentence interpreting the results in the context of the study, such as, the definition of the parameter estimated by this confidence interval, and how confident you are that the interval contains this parameter!

## Output from R

The `iscambinomtest` includes some interesting graphs.



The bottom graph illustrates the 95% confidence interval. The interval is centered at the observed sample proportion  $\hat{p}$  and then displays the two endpoints of the interval of plausible values for the process probability.

The top graph shows the null distribution assuming the lower value of the confidence interval as the process probability. This is as far left as we can shift that distribution before the area to the right of  $\hat{p}$  dips below 0.025. The middle graph shows how far we can move that distribution to the right (largest plausible value of  $\pi$ ) before the probability below  $\hat{p}$  dips below 0.025.

## Practice Problem 1.5

- Recall the 8 out of 10 statistic for St. George's Hospital (Investigation 1.3). Based on this result, what is an interval of plausible values for the underlying mortality rate at St. George's? [Hint: You can use a 95% confidence level if none is stated.] To calculate the interval, first use the [One Proportion Inference](#) applet and trial-and-error, using multiples of 0.01 for potential parameter values. Then use technology to determine the 95% confidence interval more precisely.
- Use technology to calculate the 95% confidence interval based on the 71 deaths among 361 patients. Comment on how the width and midpoint of this interval differ from the interval in (a). Explain why these changes make sense.
- Based on the interval in (b), if you were to test  $H_0: \pi = 0.20$  vs.  $H_a: \pi \neq 0.20$ , would you reject or fail to reject the null hypothesis? Explain how you know.

### Summary of Exact Binomial Inference (Sampling from a Binomial Process)

Let  $X$  represent the number of successes in the sample and  $\pi$  the probability of success for the process.

#### To test $H_0: \pi = \pi_0$

We can calculate a p-value based on the binomial distribution with parameters  $n$  and  $\pi_0$ . The p-value can be one-sided or two-sided based on the statement of the research conjecture.

if  $H_a: \pi > \pi_0$ : p-value =  $P(X \geq \text{observed})$

if  $H_a: \pi < \pi_0$ : p-value =  $P(X \leq \text{observed})$

if  $H_a: \pi \neq \pi_0$ : p-value = sum of both tail probabilities using a method like “small p-values”

#### ( $100 \times C$ )% Confidence Interval for $\pi$

The set of values such that the two-sided p-value based on the observed count is larger than the  $(1 - C)$  cut-off.

#### Technology

- **One Proportion Inference applet** for approximate and exact binomial probability (p-value)
- **R, ISCAM Workspace:** `iscambinomtest(observed, n, hypothesized π₀, alternative="greater", "less," or "two.sided", conf.level)`  
If you don't specify a hypothesized value and alternative, be sure to label the confidence level

### Handy Reminders

- **In R:** If you don't remember the inputs for an ISCAM R Workspace function, just pass “?” (with quotations) into the function, e.g., `> iscambinomtest ("?")`. When the text lists terms for the inputs (in italics), those can be used to identify the inputs and you may enter them in any order. Use the up arrow to return to earlier commands.

### Investigation 1.6: Improved Baseball Player

The previous investigation reminded you that when you fail to reject a hypothesized parameter value, you haven't *proven* that is the exact parameter value. In fact, you could be making the wrong decision. In this investigation, you will consider the types of errors that could be made in tests of significance and in the Probability Exploration you will further explore how to calculate the probabilities of these errors.

Suppose that a baseball player who has been a 0.250 career hitter suddenly improves over one winter to the point where he is now a 0.333 hitter (i.e., his probability of getting a hit has increased from 0.250 to 0.333), and so he wants a raise in salary. This may not seem like a large improvement but in major league baseball it is worth millions of dollars! However the player still needs to convince his manager that he has genuinely improved. The manager has offered to examine his performance in a certain number of at-bats (i.e., trials). The goal is to decide whether the player is being offered a fair deal.

- (a) Define the parameter of interest in words. What symbol will you use to represent it?

We know the player's probability of getting a hit is now equal to 0.333, but the manager does not know this. So, the manager will assume the player's probability is still 0.250 unless there is convincing evidence that his probability is now larger than 0.250.

- (b) State the appropriate null and alternative hypotheses that the manager wants to test. [Hint: Which is the "by chance" hypothesis and which hypothesis is the manager trying to gather evidence for?]

$$H_0:$$

$$H_a:$$

- (c) Suppose the manager decides to give the player a sample of 20 at-bats. How likely do you think it is that this 0.333 hitter will be able to convince the manager he has improved from the 0.250 hitter he used to be? Explain.

So the question is, how likely it is that someone who is now genuinely a 0.333 player will be able to demonstrate his improvement to the manager (will convince the manager to reject the null hypothesis)? We will investigate this through a two-step process. First, we need to determine how many hits the player would need to get to convince the manager that he's better than a 0.250 hitter. Then we need to determine how likely a 0.333 hitter is to perform that well.

- (d) Use the [One Proportion Inference](#) applet to see the null distribution for the number of hits obtained in 20 at-bats for a 0.250 hitter. (Check the Exact Binomial box instead of Draw Samples.) Based on your graph, (roughly) how many hits would the hitter need to get to convince you that he is better than a 0.250 hitter?

We can analyze this scenario two different ways: simulation and binomial probabilities. Let's start with simulation.

(e) Open the [Power Simulation](#) applet.

- Specify **.250** as the hypothesized probability of success
- Specify **.333** as the alternative probability of success.
- Enter **20** as the sample size
- Enter **1000** as the number of repetitions.
- Click the **Draw Samples** button.

## Power Simulation

Hypothesized probability of success:	<input type="text" value="0.25"/>
Alternative probability of success:	<input type="text" value="0.333"/>
Sample size:	<input type="text" value="20"/>
Number of samples:	<input type="text" value="1000"/>
<input type="button" value="Draw Samples"/>	

The top graph displays how many hits a 0.250 hitter got in the 20 at-bats for 1000 different sessions. The bottom graph displays how many hits a 0.333 hitter got in 20 at-bats for 1000 different sessions.

Do the two distributions have much overlap?

What does this say about how likely the player is to convince the manager that he really has improved in the sample of 20 at-bats? Explain.

Let's assume the manager will be convinced the player has improved if the player gets so many hits in his 20 at-bats, that there is less than a 5% chance that a 0.250 hitter would perform that well. In other words, the manager wants to test the above null hypothesis and needs a p-value of 0.05 or smaller in order to be convinced that the player's long-run probability of getting a hit is now above 0.250. This pre-specified cut-off value for the p-value is called the *level of significance*.

**Definition:** The [level of significance](#), often denoted by  $\alpha$  ("alpha"), is a standard that can be specified in advance for deciding when a p-value is small enough to provide convincing evidence against the null hypothesis. Then if the study's p-value  $\leq \alpha$ , we *reject* the null hypothesis in favor of the alternative and the result is said to be "statistically significant at the  $\alpha$  -level." Otherwise we "fail to reject" the null hypothesis, not finding convincing evidence to reject it. Common choices for  $\alpha$  are 0.01 and 0.05.

(f) In the applet, select **Level of Significance** from the **Choose one** pull-down menu, specify  $\alpha$  to be 0.05, and press the **Count** button. The top graph will show how many hits the player needs to get in his 20 at-bats to fall in the top 5% of the null distribution. Report this value.

**Definition:** This value designates the [rejection region](#), the values we would need to observe for the statistic in the study in order to be willing to reject the null hypothesis.

In this study, the rejection region is  $X \geq 9$ , meaning that the player needs to get at least 9 hits in his 20 at-bats in order for the p-value to be less than 0.05 and provide convincing evidence at the 5% level of significance that the player's long-run probability now exceeds 0.250.

(g) Based on the simulation results, what is the “empirical level of significance” or the estimated probability that a 0.250 hitter obtains an outcome that is in the rejection region?

**Definition:** A [Type I error](#) in a test of significance occurs when the null hypothesis is true but the test decision is to reject the null hypothesis. This type of error is sometimes referred to as a false positive or a “false alarm.” The significance level  $\alpha$  gives an upper bound on the probability of making a Type I error.

(h) Describe what Type I error means in this context.

(i) Step two is to now consider the 0.333 hitter’s point of view: What is the empirical probability that he obtains enough hits for the null hypothesis (of him being a 0.250 hitter) to be rejected? In other words, what is the empirical probability that the outcome is in the rejection region? [Hint: Think about which distribution to look at to answer this question.]

The hitter should be a bit dismayed, as the probability that he will perform well enough to convince the manager (as you calculated in h) is not all that high. He has only about a 20% chance of demonstrating his improvement through a sample of 20 at-bats.

**Definitions:**

- A Type II error occurs when we fail to reject the null hypothesis even though the null hypothesis is false. This type of error is sometimes referred to as a false negative or a “missed opportunity.”
- The [power of a test when  \$\pi = \pi\_a\$](#)  is defined as the probability of (correctly) rejecting the null hypothesis assuming this alternative value for the parameter. Thus, power reveals how likely our test is to detect a specific difference (or improvement or effect) that really is there.

Note that power =  $1 - P(\text{Type II error})$

(j) Describe what Type II error means in this context.

In this investigation, the player’s “power” is only about 0.20.

(k) *Decision Table:* The table below shows the possible states of the world and the possible decisions we can make. Indicate where the types of errors fall in this table and where the test makes the correct decision.

		State of the world	
		$H_0$ true	$H_0$ false
Test decision	Reject $H_0$		
	Fail to reject $H_0$		

(l) Which type of error (Type I or Type II) would the (improved) player prefer to have a small probability? Which type of error would the team owner prefer to have a small probability, assuming his/her priority is not to have to pay a larger salary if the player has not really improved? Explain.

Player:

Manager:

### Practice Problem 1.6A

For the research study on the mortality rate at St. George's hospital (Investigation 1.3), the goal was to compare the mortality rate of that hospital to the national benchmark of 0.15. Suppose you plan to monitor the next 20 operations, using a level of significance of 0.05. Also suppose the actual death rate at this hospital equals 0.25.

- (a) If you were to conclude that the hospital's death rate exceeds the national benchmark when it really does not, what type of error would you be committing (*Choose one*: Type I, Type II, Both, Neither)?
- (b) If you were to conclude that the hospital's death rate does not exceed the national benchmark when it really does, which type of error would you be committing (*Choose one*: Type I, Type II, Both, Neither)?
- (c) Which error, Type I or Type II, would you consider more critical here? Explain.
- (d) Use the [Power Simulation](#) applet to determine the rejection region for a sample of 20 patients.
- (e) What is the approximate power for this rejection region? Write a one-sentence interpretation of what we mean by "power" here.

## Factors that Influence Power

You have learned through this investigation that there are two types of errors that one can make with tests of significance. So far in Investigation 1.6, you found the manager might become convinced that the player has improved when he really has not (a Type I error), or the manager might be not be convinced that the player has improved when he really has (a Type II error). Now you will explore how we might reduce these error probabilities.

- (m) Suppose the manager wants to reduce the probability of a Type I error. Should he instead require the player to make at least 8 hits or at least 10 hits?

8

10

Explain.

- (n) In the **Power Simulation** applet, make the change suggested in (m) by choosing **Rejection Region** from the pull-down menu and specifying your choice for the cut-off value for the number of hits. Press **Count** and report the new empirical level of significance (estimated probability of a type I error). Did the estimated probability of a Type I error increase or decrease?

- (o) Also note the new approximate power with this smaller probability of a Type I error. Does power increase or decrease? What does this say about the probability of a Type II error?

**Discussion:** The probabilities of a Type I error and a Type II error have an inverse relationship. If we change the level of significance to decrease the probability of a Type I error, we will increase the probability of a Type II error. Power and Type II error on the other hand are directly and inversely related, as power is one minus the probability of Type II error. In some studies, you may seek a balance between these Type I and Type II errors. More typically, especially if a Type I error is considered more serious, we will fix the level of significance (controlling the probability of a Type I error), and then consider the probability of a Type II error.

- (p) What might the player ask for in order to have a better chance of showing that his success probability really has improved? Explain.

- (q) In the applet, change the **sample size** from **20** to **100** [keeping the number of samples at 1000] and press **Draw Samples**.

How did the two distributions change?

How did the overlap between the two distributions change?

What does the change in amount of overlap mean in terms of how likely the player is to convince the manager that he has improved?

(r) Use the applet to determine the rejection region corresponding to the 5% level of significance. In other words, how many hits would the 0.333 hitter need to get in 100 at-bats to convince the manager that he has improved (with a p-value  $\leq 0.05$ )?

(s) What is the approximate power of the test for a 0.333 hitter with this cut-off value? How does this approximate power compare to what you found in (i) with a sample size of 20 at-bats?

(t) So if the player really has improved, does increasing the sample size help the player? Explain why this makes intuitive sense.

(u) Now determine the rejection region so that the probability of a Type I error (level of significance) is at most 0.01 (still with a sample size of 100 at-bats).

What is the new cut-off value for the rejection region?

What is the new empirical probability of a Type II error?

Did this change help the player's likelihood of convincing the manager that he has improved? Explain.

(v) Now suppose the player actually became a 0.400 hitter, meaning that he increased his success probability to 0.4! How do you expect this to affect the power of the test (i.e., the player's ability to convince the manager he has improved)?

(w) Now specify 0.400 as the **alternative value of  $\pi$** , the probability of success, (with  $n = 100$ ), and press **Draw Samples**. How did the two distributions change? How did the overlap between the two distributions change? What is the rejection region so the probability of a Type I error is at most 0.05?

How does it compare to the first rejection region? What is the approximate power in this case? How does this probability compare to what you found in (s)?

(x) Write a paragraph summarizing your findings. Include a discussion of how the trade-off between  $P(\text{Type I error})$  and  $P(\text{Type II error})$  relates to the trade-off between making the player happy and making the manager happy with the decision making process. [*Hint:* What happens if we try to decrease the probability of one of the errors?] Also discuss how these errors are affected by changes in sample size, level of significance, and the alternative probability of success (the player's true improved performance).

**Discussion:** You have learned through this investigation that there are two types of errors that one can make with tests of significance. In this example the manager might become convinced that the player has improved when he really has not (a Type I error), or the manager might not be convinced that the player has improved when he really has (a Type II error). Naturally, we would like the probabilities of both types of errors to be small, but unfortunately they are inversely related: as one error probability decreases, the other increases (unless other factors change also). What's typically done in practice is to set the maximum allowable probability of Type I error in advance by setting the level of significance  $\alpha$ , the most common value is 0.05, followed by 0.10 and 0.01, and then determine the sample size necessary for the probability of a Type II error to be below a specific value.

You have also learned in this investigation that increasing the sample size can decrease both error probabilities, but the downside is that using a larger sample size requires more time, effort, and expense. Or, if you keep the probability of Type I error fixed, increasing the sample size decreases the probability of Type II error. So, larger samples are better, but the trade-off is that they typically come at a price of time, or money, or both. Another way that the probability of Type II error will be smaller, while keeping the Type I error probability fixed, is if the difference between the hypothesized and the actual success probability is larger. Although that is not something you can control.

Power is closely related to the probability of Type II error, because  $\text{power} = 1 - P(\text{Type II error})$ . Power is the probability of correctly rejecting a null hypothesis that is false, so we prefer statistical tests with large power. You have learned that power is influenced by the following factors:

- Increasing sample size increases power.
- Increasing the significance level  $\alpha$  increases power.
- Increasing the distance between the alternative and null values increases power.

### Practice Problem 1.6B

Again consider the research study at St. George's hospital (Investigation 1.3), suppose the probability of a death during transplantation at St. George's was 0.20.

- In a sample size of 50 operations, what is the probability that we will reject the null hypothesis that the rate at St. Georges equals 0.15 in favor of the alternative hypothesis that the rate exceeds 0.15, with a 5% level of significance?
- How does your result compare to Practice Problem 1.6A, where you looked at an alternative probability of 0.25 and a sample size of 20. Does this comparison match what you would predict? (Discuss all relevant factors, even if they might contradict each other.)
- If Abe were to use a significance level of 0.05 and Bianca were to use a significance level of 0.01, who would have a smaller probability of Type I error? Explain briefly.
- If Abe were to use a significance level of 0.05 and Bianca were to use a significance level of 0.01, who would have a smaller probability of Type II error? Explain briefly.

## Probability Exploration: Exact Binomial Power Calculations

We can use technology to obtain the rejection region and binomial probabilities of Type I and Type II errors based on the binomial distribution. The calculation of power will be a two-step process.

- *Step one:* Determine the rejection region corresponding to the null hypothesis hypothesized value, the direction of the alternative hypothesis, and the level of significance.
- *Step two:* Determine the probability of obtaining an observation in the rejection region for a specific alternative value of the parameter.

To determine the rejection region we work “in reverse”: specifying a probability and asking the software to determine the corresponding number of successes.

### Technology Detour – Determine the Rejection Region (Binomial)

#### In R

- The `iscaminvbinom` function (of the ISCAM workspace) takes the following input
  - *alpha* = the probability of interest (the level of significance)
  - *n* = the sample size (number of trials)
  - *prob* = the process probability ( $\pi$ ) of the binomial distribution
  - *lower.tail* = TRUE or FALSE

For example: > `iscaminvbinom(alpha=.05, n=20, prob=.25, lower.tail=FALSE)` should reveal a graph with the upper 5% of the distribution shaded in red, and  $X = \dots$  for the smallest cut-off value that has at most 0.05 probability above it (the actual probability above that cut-off value will also be displayed).

- (a) Based on the output from the technology, what is the smallest number of successes  $k$  such that  $P(X \geq k) \leq 0.05$  when  $n = 20$  and  $\pi = 0.25$ ? How does this compare to the simulation results? What is the exact probability of a Type I error in this case?

Now that we have the rejection region, we switch to the alternative probability of success and see how often we land in that rejection region.

### Technology Detour – Calculating Power from Rejection Region (Binomial)

#### In R

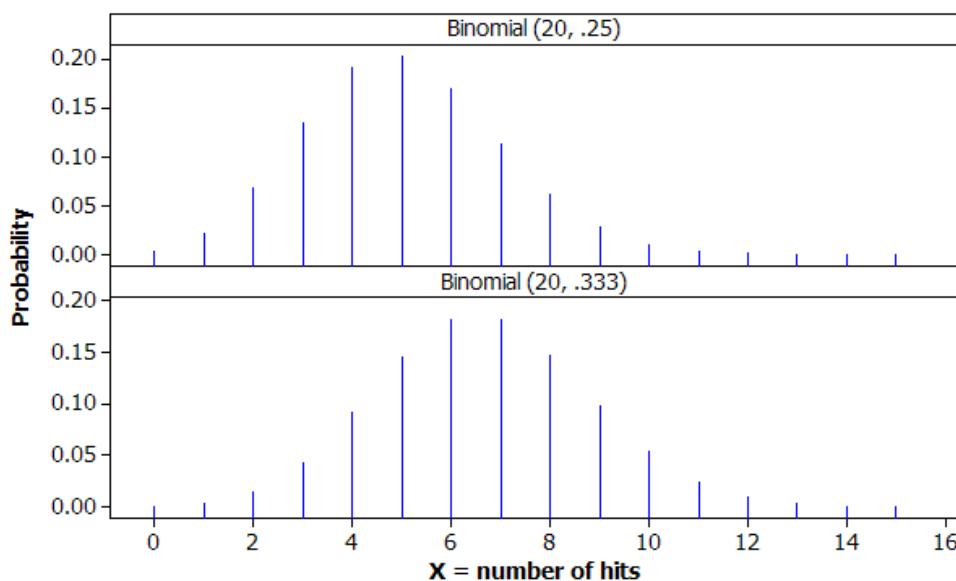
- Use `iscambinomprob` as before, specify 0.333 as the probability of success, but staying in upper tail to match our alternative hypothesis.  

```
> iscambinomprob(k=9, n=20, prob=.333, lower.tail=FALSE)
```

- (b) What is the exact probability that a 0.333 hitter will have at least 9 hits (as defined by the rejection region in (a)) in 20 at-bats? That is, how often will we reject (at the 5% level) the null hypothesis that  $\pi = 0.250$  when  $\pi$  actually equals 0.333? How does this compare to the approximate probability given by the applet? In addition to calculating this probability, fill in the following blanks to indicate how you calculate it.

Power =  $P(Y \geq \underline{\quad})$ , where Y has a Binomial distribution with  $n = \underline{\quad}$  and  $\pi = \underline{\quad}$   
 $= \underline{\quad}$

The following graphs are the binomial distributions for samples of size  $n = 20$  with  $\pi = 0.250$  and  $\pi = 0.333$  on the same scale.



(c) Circle on these graphs which vertical bars correspond to the probability of a Type I error, which the probability of a Type II error, and which the power of the test. Make sure you clarify which distribution you are using.

### Sample Size Determination

- (d) Suppose the player wants the power of correctly determining that he has become a 0.333 hitter to be at least 0.80. Use trial-and-error to determine the sample size necessary for him to achieve this.  
 (Continue to assume the manager uses 0.05 as the level of significance.) [Hint: Remember that you need to redo both steps in this process! Don't forget to switch the value of  $\pi$  ...]

### Practice Problem 1.6C

(a) Repeat (a)-(d) of the Exploration using 0.01 as the level of significance.

(b) Did the sample size necessary to achieve power of at least 0.80 become larger or smaller with this smaller significance level? Explain why this makes intuitive sense.

## SECTION 2: NORMAL APPROXIMATIONS FOR SAMPLE PROPORTIONS

So far, all the studies you have examined have involved one categorical binary variable and the relevant statistic has been the number of successes, or equivalently the proportion of successes, where “success” is defined as the outcome you count (e.g., number of heart transplantation deaths, number of helper toys, number of hits) in a particular sample. You have used simulation from a coin-tossing type process to approximate probabilities (p-value, error probabilities) and you calculated exact probabilities using the binomial distribution.

Historically, when computers were less prevalent, these empirical and binomial probabilities were often cumbersome to determine and statisticians instead applied a far more convenient approximation method to estimate the probabilities of interest. In this section, you will see how the normal probability model can be used as a third method to approximate the p-values, confidence intervals, and error probabilities from the previous section. See Chapter 2 for more information about the normal distribution.

### Investigation 1.7: Reese’s Pieces

Manufacturers often perform “quality checks” to insure that their manufacturing process is operating under certain specifications. Suppose a manager at Hershey’s is concerned that his manufacturing process is no longer producing the correct color proportions in Reese’s candies.

- (a) Take a (presumably representative) sample of  $n = 25$  Reese’s Pieces candies and record the number of orange, yellow, and brown candies in your sample.

Orange:

Yellow:

Brown:

- (b) Identify the observational units and variable for your sample.

Observational units:

Variable:

Type:

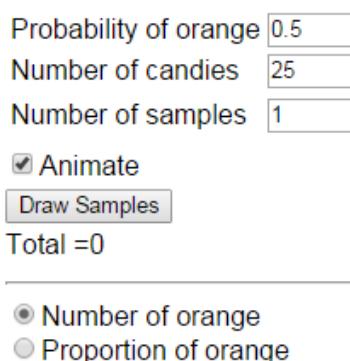
- (c) Define “success” to be an orange candy and “failure” to be a non-orange candy. Is it reasonable to treat these data as observations from a binomial process?

- (d) Based on your data, do you think 0.50 is a plausible value for the probability Hershey’s process produces an orange candy? How are you deciding?

(e) Open the [Reese's Pieces](#) applet. Note that we have set the applet to assume a 0.50 process probability of a Reese's Pieces candy being orange. The applet is also set to take a sample of  $n = 25$  candies to match your class investigation.

- Click the **Draw Samples** button. The applet will randomly select a sample of 25 candies, sort them, and report the sample number of orange candies.
- Click the **Draw Samples** button again.

Did you get the same number of orange candies both times?



*Note:* The applet is assuming a binomial process with  $\pi$  equal to the probability of orange that you specify. We see that sampling variability is alive and well.

- (f) Change the **Number of samples** from 1 to 998, uncheck the **Animate** box, and click the **Draw Samples** button again. Check the **Summary Stats** box. Describe the behavior (shape, center, variability) of the resulting null distribution of these sample counts. Where does your observed number of orange candies from (a) fall in this distribution?

- (g) Now select the *Proportion of orange* radio button. How does this change the null distribution? How does this change the rareness of your observed result?

**Discussion:** Rescaling the simulated statistics by dividing by the sample size does not change the shape of the distribution or the relative ordering of the observations, but will make it easier to compare distributions as all values must fall between 0 and 1, inclusive. But can we easily predict the center and variability?

- (h) If we had instead taken 1000 samples of size  $n = 100$  candies, how do you think the distribution of the sample proportions would compare to the distribution where  $n = 25$ ? Explain.

- (i) Without pressing Reset, change the **Number of candies** in the applet to **100**, the **Number of samples** to 1000, and press **Draw Samples** to generate a new distribution. Check the **Show Previous** box to display the previous distribution in the background in light grey. Describe the behavior (shape, center,

variability) of this (new) distribution and how it has changed from the previous. Focus on the most substantial change in the distribution. Is this what you expected?

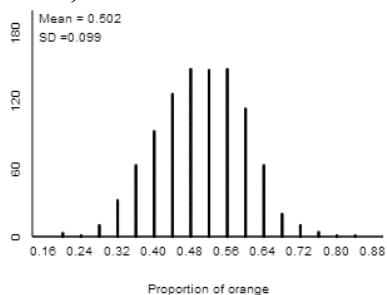
(j) Now let's suppose the manufacturing process was set to produce 25% orange candies in the long-run. Model this by assuming the value of  $\pi$  to be 0.25. How do you anticipate this will change the distribution of sample proportions? Explain.

(k) Create and describe the distribution of sample proportions in this case ( $\pi = 0.25, n = 100$ ). What is the primary difference in how the distribution of sample proportions of successes has changed with this change in the process probability?

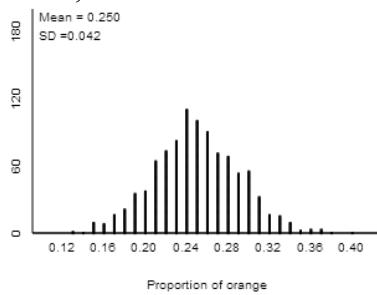
(l) Suppose we think the probability of being orange is 0.8. Repeat the previous question with  $\pi = 0.8, n = 100$ .

Below are simulation results for the three scenarios you have examined:

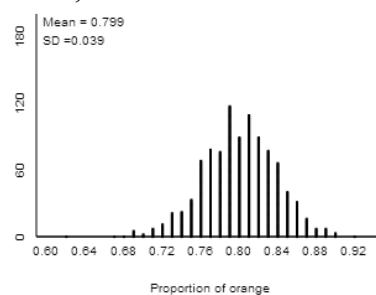
$n = 25, \pi = 0.5$



$n = 100, \pi = 0.25$



$n = 100, \pi = 0.80$



(m) Summarize the differences you have noted between these distributions and also what characteristic(s) they all have in common.

Differences:

Similarities:

A very common and useful probability model is the *normal distribution*.

### Probability Detour – Normal Random Variables

Many quantitative variables have a distribution that can be reasonably modeled with a [normal probability curve](#). This is a continuous probability distribution (as opposed to the binomial, which is a discrete distribution) that is specified by two quantities:

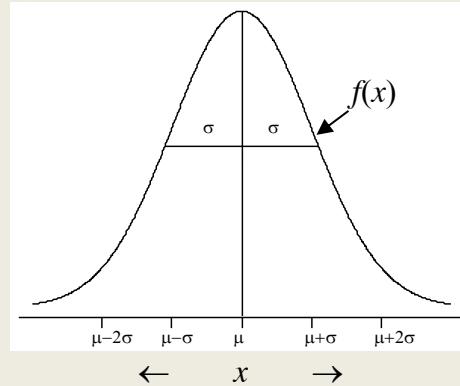
- Mean, denoted by  $\mu$ , which is the peak and point of symmetry
  - Standard deviation, denoted by  $\sigma$ , which is the distance between the mean and the inflection points
- In other words, the mean determines the center of the distribution, and the standard deviation controls how variable (spread out) the distribution is. Standard deviation can be loosely interpreted as a typical distance of the observations from the mean.

This normal curve is given by the following function:

$N(\mu, \sigma)$ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty < x < \infty$$

Notes: The symbol  $\pi$  is used here to refer to the irrational number  $\approx 3.14159\dots$ . The constant in front of the exponential term ensures that the total area under any normal curve equals one.



Then the probability of an observation falling in any particular interval is determined by the area under the normal curve in that interval. In principle, this area could be determined by integrating the function  $f(x)$  above. But this particular function has no closed form anti-derivative, so we rely on technology to approximate these areas and therefore these normal probabilities.

For now, we will focus on using this mathematical model as an approximation to the distribution of sample proportions. In this case, one can show (using some basic probability rules) that the **theoretical mean** of the distribution of sample proportions will be equal to the process probability  $\pi$  and the

**theoretical standard deviation** of this distribution equals  $SD(\hat{p}) = \sqrt{\frac{\pi(1-\pi)}{n}}$ .

(n) Calculate the theoretical mean and standard deviation of the sampling distribution of sample proportions for each of these three cases.

	Theoretical mean of $\hat{p}$	Theoretical SD of $\hat{p}$
$n = 25, \pi = 0.5$		
$n = 100, \pi = 0.25$		
$n = 100, \pi = 0.80$		

How do the theoretical means and standard deviations compare to the simulated values (see the graphs before part (m)) and to each other?

- (o) Now suppose we had taken samples of size  $n = 5$  candies instead (from a process with success probability 0.80). Predict how the distribution of sample proportions will behave (shape, center, variability).

Shape:

Mean:

SD:

- (p) Use the applet to check your predictions. Discuss your observations.

**Discussion:** It is very important to keep in mind that this normal probability model is not always a valid approximation for the distribution of sample proportions. Whether it is valid will be determined by a combination of the sample size  $n$  (larger samples result in more symmetric distributions) and the value of  $\pi$  (values closer to 0 or 1 result in less symmetric distributions). A common guideline is to assume symmetry when  $n \times \pi \geq 10$  and  $n \times (1 - \pi) \geq 10$ .

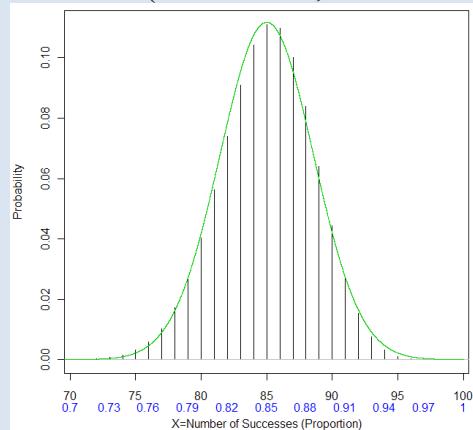
- (q) Explain how these guidelines are consistent with your observations above.

This result about the long-run pattern of variation in sample proportions is more formally called the [Central Limit Theorem](#) (CLT) and is one of the most important results in Statistics. It states that if the sample size is large enough, then the *sampling distribution* of the sample proportion  $\hat{p}$  will be well modeled by a normal distribution with mean  $E(\hat{p})$  equal to  $\pi$ , the process probability, and standard deviation

$$\text{SD}(\hat{p}) \text{ equal to } \sqrt{\frac{\pi(1-\pi)}{n}}.$$

The convention is to consider the sample size large enough if  $n \times \pi \geq 10$  and  $n \times (1 - \pi) \geq 10$ .

Binomial( $n = 100, \pi = 0.80$ ) / 100  
vs. Normal (mean = 0.80, SD = 0.040)



**Discussion:** Use of the normal probability model to approximate the sampling distribution of sample proportions is largely historical (back when they didn't have computers, the binomial distribution calculations were a bit of a pain but much easier with the normal distribution). But you will see in the next few investigations that some calculations are still more convenient with the normal distribution.

For example, we stated in Investigation A that we usually begin to think an observation is unusual when it lies more than two standard deviations above or below the mean of the distribution. This comes from a very special property of the normal probability model.

(r) Reset the Reese's Pieces applet, specify  $\pi = 0.25$  and  $n = 100$ , and check the **Exact Binomial** and **Normal Approximation** boxes. Now use the theoretical mean and SD values from (n) to calculate the value of  $\hat{p}$  that is two standard deviations below the mean. Enter this value in the **As extreme as** box, toggle to “less than” and press Count.

Mean:

SD:

Mean – 2SD:

Probability below  
Binomial model:

Normal model:

(s) How do the probabilities compare?

(t) Write a one-sentence interpretation of this probability.

(u) Now check the **Two-tailed** box. What is the two-tailed probability for the normal distribution?

(v) Explain how your observation in (u) confirms the statement we made in Investigation A. [*Hint:* How often do we obtain an observation more than 2SDs from the mean in a normal distribution?]

(w) Now check the **between** box. Write a one-sentence interpretation of the 0.9545 probability.

Roughly 95% of ....

### Practice Problem 1.7

(a) For which of the earlier studies we have investigated would a normal approximation be appropriate? Explain.

- Friend or Foe
- Do Names Match Faces
- Heart Transplant Mortality (10 cases)
- Heart Transplant Mortality (361 patients)
- Kissing the Right Way

(b) In this investigation, you took samples of 25 candies. Between what two values should 95% of the sample proportions fall when  $\pi = 0.50$ ?

### Investigation 1.8: Is ESP real?

Statistician Jessica Utts has conducted extensive analysis of studies that have investigated psychic functioning. (Combining results across multiple studies, often to increase power, is called *meta-analysis*.) Utts (1995) cites research from Bern and Honorton (1994) that analyzed studies that used a technique called ganzfeld.

In a typical ganzfeld experiment, a “receiver” is placed in a room relaxing in a comfortable chair with halved ping-pong balls over the eyes, having a red light shone on them. The receiver also wears a set of headphones through which [static] noise is played. The receiver is in this state of mild sensory deprivation for half an hour. During this time, a “sender” observes a randomly chosen target and tries to mentally send this information to the receiver ... The receiver is taken out of the ganzfeld state and given a set of possible targets, from which they must decide which one most resembled the images they witnessed. Most commonly there are three decoys along with a copy of the target itself. [[Wikipedia](#)]

- (a) Suppose you want to test whether the subjects in these studies have ESP, with  $\pi$  equal to the actual probability that they identify the correct image. State appropriate null and alternative hypotheses, in symbols and in words.

$$H_0:$$

$$H_a:$$

Utts cites Bern and Honorton as reporting that the studies showed 106 “hits” in 329 sessions. (Dr. Utts’ report: [www.ics.uci.edu/~jutts/air.pdf](http://www.ics.uci.edu/~jutts/air.pdf).)

- (b) Determine the sample proportion of hits, and denote it with an appropriate symbol.

- (c) Although we know how to compute the exact binomial p-value for this test, in this investigation you will use the normal probability model to approximate the binomial calculation. Assuming that the subjects have no psychic ability, is this sample size large enough to allow using the Central Limit Theorem? Justify your answer with appropriate calculations.

- (d) Describe what the Central Limit Theorem says about the distribution of sampling proportions in this context, assuming the null hypothesis is true, and produce a well-labeled sketch to illustrate your description.

- (e) In your graph, shade the area under the curve corresponding to the sample proportion of “hits” being 0.322 or higher.

The area you have shaded represents the probability of a sample proportion exceeding 0.322, under the assumption that the subjects have no psychic ability (with  $n = 329$  sessions).

(f) Based on your shading, provide a rough guess of this area as a percentage of the total area under this normal curve.

(g) How many standard deviations (SDs) above the mean is the value 0.322, the observed sample proportion of hits? [Hint: First subtract the mean from 0.322, then divide the result by the SD.]

**Definition:** The standardized score of an observation determines the number of standard deviations between the observation and the mean of the distribution:

$$z = \frac{\text{observation} - \mu}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

This quantity is also referred to as the  $z$ -score. By converting to this  $z$ -score, we say we have *standardized* the observation. This provides us another metric or ruler for how unusual an observation is. The sign of the  $z$ -score tells us whether the observation falls above or below the mean. If  $X$  follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then  $Z$  follows a normal distribution with mean 0 and standard deviation 1.

Because this  $z$ -score is larger than 2, we already know the probability to the right is rather small. But how can we determine this probability more precisely? Generally we would integrate the function over the region of interval  $(0.322, \infty)$ . Because that is not possible with the normal probability function, we will instead use technology, which implements numerical integration techniques to calculate this area/probability.

### Technology Detour – Calculating Normal Probabilities

**In R:** The `iscamnormprob` function takes the following inputs:

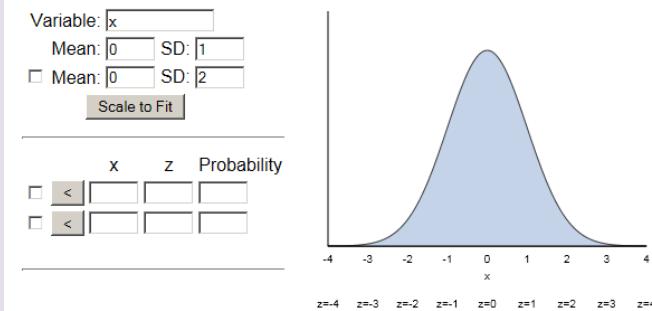
- `xval` = the  $x$  value of interest
- `mean` = the mean of the normal distribution
- `sd` = the standard deviation of the normal distribution
- `direction` = “above,” “below,” “outside” or “between” (using the quotes)
- `label` = a string of text (in quotes) to put on the horizontal axis
- `xval2` = an optional input for use with “outside” and “between” directions.
- `digits` = specifies how many significant digits to display, default is 4

For example (with or without input labels): `iscamnormprob(xval=.322, mean=.25, sd=.02387, direction="above", label="sample proportions")`

[Normal Probability Calculator Applet](#)

- Type in a description of the variable in the **variable** box. (Here, the variable is the sample proportion.)
- Specify the values of the theoretical **mean** and **standard deviation** for the  $\hat{p}$  distribution.
- Press the **Scale To Fit** button to redraw and scale the curve.
- Check** the box next to the first < sign to activate that row.
- Specify the value of interest in the X box. Press **Enter/Return**.

You can also check the Normal Approximation button in the One Proportion applet.

**Normal Probability Calculator**

(h) Use technology (for the normal probability model specified by the CLT) to approximate the probability that the sample proportion of “hits” would exceed 0.322, still assuming that the subjects have no psychic ability. How does this value compare to your guess in (f)?

(i) Compute the exact binomial p-value. How does the normal probability model calculation compare to the exact binomial calculation? Does the normal model provide a close approximation to the exact p-value in this case?

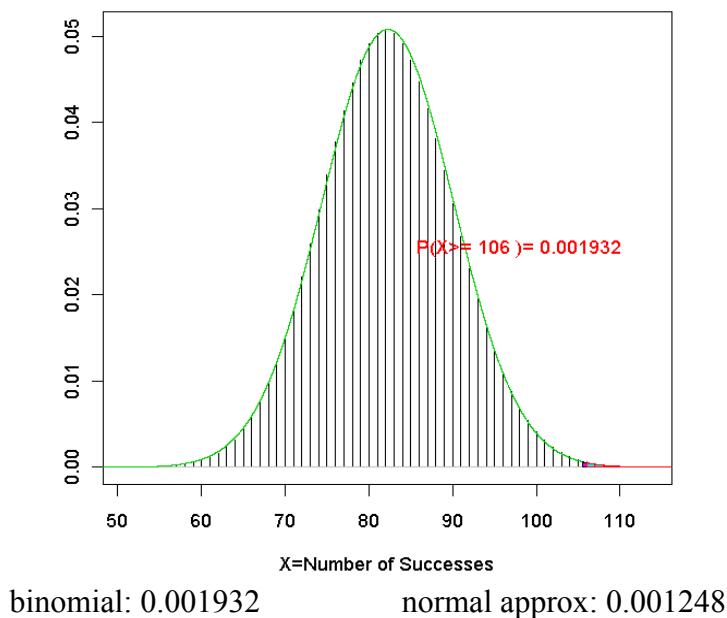
(j) Provide a one-sentence interpretation of the probability you have calculated,  $P(X \geq 106) \approx P(\hat{p} \geq 0.322)$ . What conclusion would you draw from this probability calculation? Explain.

Interpretation:

Conclusion:

### Study Conclusions

The data reported by Utts corresponds to a sample proportion  $\hat{p}$  equal to 0.322. This is clearly larger than the 0.250 we would expect if the subjects were simply guessing. Could this higher sample proportion have resulted from chance variability from a process with  $\pi = 0.250$ ? Applying the Central Limit Theorem (which is valid because  $n = 329$  so  $n \times \pi = 329(0.25) = 82.25 > 10$  and  $n \times (1-\pi) = 329 \times 0.75 = 246.75 > 10$ ) we approximate the distribution of sample proportions (assuming  $\pi = 0.25$ ) with a normal curve with mean 0.25 and standard deviation equal to  $\sqrt{\pi(1-\pi)/n} = \sqrt{.25(.75)/329} = 0.02387$ . So we have some information about the unusualness of this observation when we determine 0.322 is more than 3 standard deviations from the hypothesized value of 0.25. Using technology, we approximate  $P(\hat{p} \geq 0.322) \approx P(Z \geq 3.02) \approx 0.0012$ . We could get a sample proportion this extreme by chance alone if the subjects were all guessing, but it would be highly unlikely. This provides very strong evidence against the null hypothesis that such subjects are guessing. If we find  $P(X \geq 106)$  exactly from a Binomial distribution with  $n = 329$  and  $\pi = 0.25$ , this probability equals 0.0019, so there is considerable agreement in the areas under the curve with the normal probability curve and the exact binomial probability distribution in this case (see figure below).



### Practice Problem 1.8

- Use the normal probability model to approximate the probability that less than 22% of the 329 sessions would result in “hits,” assuming the null hypothesis is true. [Start with a well-labeled sketch, shading the area of interest. Make a guess as to the value of that area and then use technology to carry out the calculation.]
- Repeat (a) to approximate the probability that between 22.6% and 27.4% of the 329 sessions would result in “hits.” Do you notice anything familiar about this probability?
- Determine how many standard deviations above the mean (assuming no psychic ability) the value  $\hat{p} = 0.274$  is. Repeat for  $\hat{p} = 0.322$ . What about 0.226?

0.274:

0.322:

0.226:

**Investigation 1.9: Halloween Treat Choices**

Obesity has become a widespread health concern, especially in children. Researchers believe that giving children easy access to food increases their likelihood of eating extra calories. Schwartz, Chen, and Brownell (2003) examined whether children would be willing to take a small toy instead of candy at Halloween. They had seven homes in 5 different towns in Connecticut present children with a plate of 4 toys (stretch pumpkin men, large glow-in-the-dark insects, Halloween theme stickers, and Halloween theme pencils) and a plate of 4 different name brand candies (lollipops, fruit-flavored chewy candies, fruit-flavored crunchy wafers, and “sweet and tart” hard candies) to see whether children were more likely to choose the candy or the toy. The houses alternated whether the toys were on the left or on the right. Data were recorded for 284 children between the ages of 3 and 14 (who did not ask for both types of treats).

- (a) Identify the observational units and variable of interest in this study. Is this variable quantitative or categorical? If categorical, how many different outcomes are possible?

Observational units:

Variable of interest:

Variable type:

Possible outcomes:

- (b) Define (in words) the parameter of interest in this study.

- (c) State an appropriate null and alternative hypothesis involving this parameter (in symbols and in words), for testing whether there is strong evidence of a preference for either the toys or the candy.

$H_0$ :

$H_a$ :

- (d) In the sample of 284 children, 135 children chose the toy and 149 chose the candy. Calculate the value of a relevant statistic in this study.

- (e) Is the sample size large enough for the conditions of the Central Limit Theorem to be met? Justify your answer with appropriate calculation(s).

(f) According to the Central Limit Theorem, in considering the distribution of sample proportions

- What is the shape of the null distribution here?
- What is the mean of this distribution?
- What is the standard deviation of this distribution?
- Create a well-labeled sketch of the distribution under the null hypothesis; shade the region(s) of interest to represent the two-sided p-value for this test; and conjecture the magnitude of the p-value based on your sketch.

(g) Calculate the  $z$ -score for the observed sample proportion compared to the probability hypothesized by the null hypothesis.

(h) What does it mean for this  $z$ -score to be negative? What do you know about the  $z$ -score for the upper tail of the two-sided p-value? Explain.

**Definition:** The above calculation result comparing the observed statistic to the hypothesized parameter value is more generally referred to as the [test statistic](#). In the case of a single categorical variable with a process probability of  $\pi_0$  (the value specified by the null hypothesis), the formula is:

$$z_0 = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

This test statistic is simply the application of the z-score from the previous investigation, and you will interpret it as “how many standard deviations the observed proportion  $\hat{p}$  lies from the hypothesized process proportion.” With the two-sided p-value, we are interested in how often we will obtain a  $\hat{p}$  value as far from the hypothesized probability *in either direction*. Due to the symmetry of the normal curve, this only involves doubling one of the tail probabilities.

- (i) Based on the test statistic value you calculated, does the sample proportion appear to be extreme under the null hypothesis?
- (j) Use technology (e.g., [Normal Probability Calculator](#) applet) to determine the approximation to the two-sided p-value with the normal distribution. Is your p-value consistent with your sketch in (f)?
- (k) Provide a one-sentence interpretation of this p-value.
- (l) Based on this p-value, will you reject the null hypothesis at the 5% level of significance?
- (m) Do you think the researchers are pleased by the lack of significance in this test? Explain, in the context of this study, why such a result might be good news for them.
- (n) Does this mean the researchers can conclude that they have proven that children do not have a preference between candy and toys? Explain.

## Study Conclusions

If we define  $\pi$  to be the probability that, when presented with a choice of candy or a toy while trick-or-treating, a child chooses the toy, and if we assume the null hypothesis ( $H_0: \pi = 0.5$ ) is true, the above calculations tell us that we would observe at most 135 children choosing candy (at least 149 choosing toy) or at most 135 of the 284 children choosing the toy in about 40% of samples. Thus, this is not a surprising outcome when  $\pi = 0.5$ . We fail to reject the null hypothesis and conclude that it's plausible that children are equally split in preferring the toy or the candy.

We do have some cautions with this study as it was conducted in only a few households in Connecticut, a "convenience sample," so we cannot claim that these results are representative of children in other neighborhoods. We also don't know if the children found the toys "novel" and whether their preference for toys could decrease as the novelty wears off (or if "better" candy choices were offered). Furthermore, when the children approached the door they were asked their age and gender, and for a description of their Halloween costume. The researchers caution that that this may have cued the children that their behavior was being observed (even though their responses were recorded by another research member who was out of sight) or that they should behave a certain way. Still, these researchers were optimistic that alternatives could be presented to children, even at Halloween, to lessen their exposure to large amounts of candy.

## Improving the approximation

- (o) Use technology to determine the exact binomial p-value. How does it compare to the normal approximation in this case?

Even though our sample size is reasonably large (for the Central Limit Theorem), there is still some "discreteness" in the exact binomial distribution that makes the normal approximation (which, of course, is continuous) less accurate than we would like. With the normal distribution,  $P(X \leq k) = P(X < k)$  because  $P(X = k) = 0$  for any particular (integer) value  $k$ . However,  $P(X = k) > 0$  with the binomial distribution. It is the failure to include the "mass" at  $k$  that can lead the normal approximation to be poor. How can we correct for this? We can apply a **continuity correction** by using the normal distribution to calculate  $P(X \leq k + 0.5)$  as an approximation for  $P(X \leq k)$  and  $P(X \geq k - 0.5)$  as an approximation for  $P(X \geq k)$ .

- (p) In order to calculate  $P(X \leq 135)$ , what probability does the continuity correction tell us to calculate?
- (q) In order to calculate  $P(X \geq 149)$ , what probability does the continuity correction tell us to calculate?  
*[Hint: Remember what the continuity correction is trying to accomplish.]*

- (r) Use technology (e.g., [Normal Probability Calculator](#) applet) to re-approximate the p-value using the normal probability distribution with the continuity correction. How does this new approximation compare to the earlier one (without the continuity correction) and to the exact binomial p-value?

Many software programs allow you to do this continuity correction for a one-sample  $z$ -test (e.g., using  $\hat{p} = (X \pm 0.5)/n$  as the input) or may do so by default. However, when  $n$  is large, you may not see much difference in the values.

### Technology Detour – Normal Approximation to Binomial

#### Theory-Based Inference Applet

- Enter the sample size and either the count or the proportion of successes, press **Calculate** to display sample data.
- Check the box for **Test of significance**
- Specify the hypothesized value
- Use the button to specify the direction of the alternative
- Press the **Calculate** button.
- Check the **cont corr.** box to apply the continuity correction to the test statistic and p-value.

Scenario: One proportion

Paste Data

n:

count:

sample  $\hat{p}$ :

**Calculate**

**Theory-Based Inference**

Test of significance

$H_0: \pi = 0.5$

$H_a: \pi < 0.5$

**Calculate**

standardized statistic   cont corr.

p-value

In R: The `iscbinomnorm(k, n, prob, direction)` function gives a visual of this continuity correction. It takes the following inputs:

- $k$  = the observation of interest
- $n$  = the sample size
- $prob$  = the process probability ( $\pi$ )
- $direction$  = “below,” “above,” or “two.sided”

- (s) Use technology to verify your calculations.

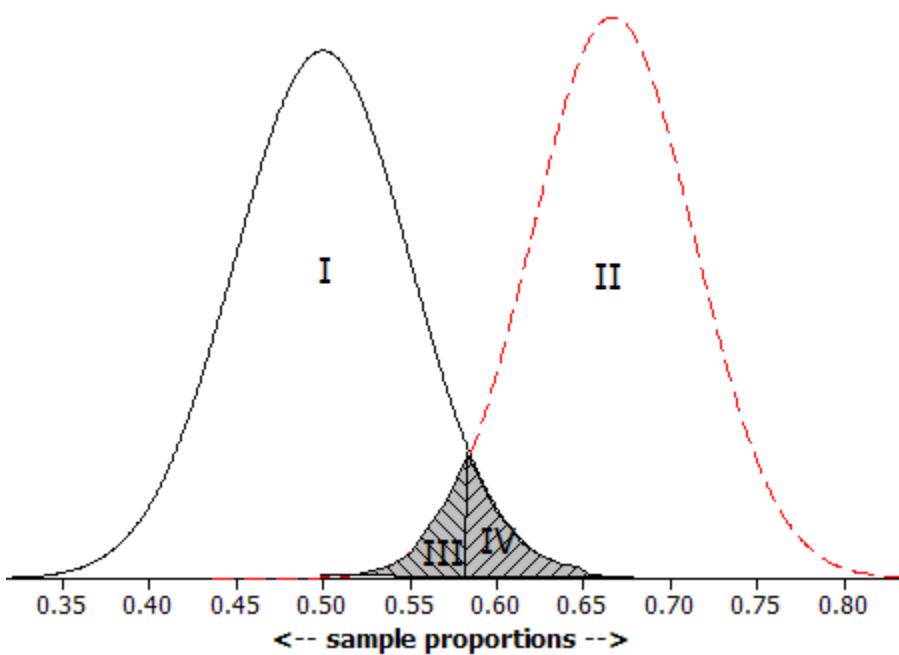
**Practice Problem 1.9A**

A student wanted to assess whether her dog Muffin tends to chase her blue ball and her red ball equally often when they are rolled at the same time. The student rolled both balls a total of 96 times, each time keeping track of which ball Muffin chased. The student found that Muffin chased the blue ball 52 times and the red ball 44 times. Let's treat the blue ball as "success."

- State the appropriate null and alternative hypotheses in symbols. (Be sure to describe what the symbol represents in this context.)
- Report and interpret the values of the  $z$ -test statistic and normal-based p-value.
- Summarize your conclusion about the student's question concerning her dog Muffin.
- Explain the reasoning process behind your conclusion.
- Suppose you want to redo this study using a sample size of 200 tosses. You plan to use a significance level of  $\alpha = 0.05$ , and you are concerned about the power of your test when  $\pi = 0.60$ . Explain how you would determine this power (the two steps) and interpret what "power" implies in this context.
- How would your answers to the previous questions change if we had used the red ball as success?

**Practice Problem 1.9B**

Suppose that you want to re-conduct the kissing study in a large city with a sample of 100 kissing couples. You want to test the null hypothesis  $H_0: \pi = 0.667$  against a one-sided alternative  $H_a: \pi < 0.667$  using a significance level of  $\alpha = 0.05$ , and you are concerned about the power of your test when  $\pi = 0.5$ . Consider the following graph:



- (a) Which region(s) represents the probability of making a Type I error?

I                  II                  III                  IV

- (b) Which region(s) represents the probability of making a Type II error?

I                  II                  III                  IV

- (c) Which region(s) represents the power of the test?

I                  II                  III                  IV

## Summary of one proportion z-test

Let  $\pi_0$  represent the (constant) probability of success for the process assumed by the null hypothesis.

### To test $H_0: \pi = \pi_0$

We can calculate a p-value based on the normal distribution with mean equal to  $\pi_0$  and standard deviation equal to  $\sqrt{\pi_0(1 - \pi_0)/n}$

**Test Statistic:**  $z_0 = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$  which follows a  $N(0,1)$  distribution

#### p-value:

if  $H_a: \pi > \pi_0$ : p-value =  $P(Z \geq z_0)$

if  $H_a: \pi < \pi_0$ : p-value =  $P(Z \leq z_0)$

if  $H_a: \pi \neq \pi_0$ : p-value =  $2P(Z \geq |z_0|)$

**Validity:** This procedure is considered valid as long as the sample size is large relative to the hypothesized probability ( $n \times \pi_0 \geq 10$  and  $n \times (1 - \pi_0) \geq 10$ ), and you have a representative sample from the process of interest.

## Technology

- **Theory-Based Inference applet**

Specify the hypothesized value, use the button to specify the direction of the alternative (< > for not equal), enter the sample size and either the count or the proportion of successes. Press the Calculate button.

- **R, ISCAM Workspace:** `iscamonepropztest(observed, n, hypothesized pi_0, alternative="greater", "less", or "two.sided", conf.level)`

You can enter either the number of successes or the proportion of successes ( $\hat{p}$ ) for the “observed” value. If you don’t specify a hypothesized value and alternative, be sure to label the confidence level

**Investigation 1.10: Kissing the Right Way (cont.)**

Recall the study published in *Nature* that found 64.5% of 124 kissing couples leaned right to kiss (Investigations 1.4 and 1.5). We previously used simulation and the binomial distribution to determine which values were plausible for the underlying probability that a kissing couple leans right. In particular, we found 0.5 and 0.74 were not plausible values but 0.6667 was. Now you will consider applying the normal model as another method for producing confidence intervals for this parameter.

- (a) With a sample size of 124 kissing couples, does the Central Limit Theorem predict the normal probability distribution will be a reasonable model for the distribution of the sample proportion?

*Note:* When you do not have a particular value to be tested for the process probability, it's reasonable to use the sample proportion in checking the sample size condition for the CLT. (Note: This is equivalent to making sure there are at least 10 successes and at least 10 failures in the sample.)

- (b) Do you have enough information to describe and sketch the distribution of the sample proportion as predicted by the Central Limit Theorem? Explain.

- (c) Suggest one method for estimating the standard deviation of this distribution of sample proportions based on the observed sample data.

**Definition:** The standard error of the sample proportion,  $SE(\hat{p})$ , is an estimate for the standard deviation of  $\hat{p}$  (i.e.,  $SD(\hat{p})$ ) based on the sample data, found by substituting the sample proportion for  $\pi$ :  $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ .

- (d) Now consider calculating a 95% confidence interval for the process probability  $\pi$  based on the observed sample proportion  $\hat{p}$ . Calculate the standard error of  $\hat{p}$ . Then, how far do you expect to see a sample proportion fall from the underlying process probability? [Hint: Assuming a normal distribution ... 95% .... ]

$$SE(\hat{p}) =$$

$$\text{Plausible distance} =$$

- (e) Use the distance in (d) and the observed sample proportion of  $\hat{p} = 0.645$  to determine an interval of plausible values for  $\pi$ , the probability that a kissing couple leans to the right.

An approximate 95% confidence interval for the process probability based on the normal distribution would be  $\hat{p} \pm 2\sqrt{\hat{p}(1 - \hat{p})/n}$ . That is, this interval extends two standard deviations on each side of the sample proportion. We know that for the Normal distribution roughly 95% of observations (here sample proportions) fall within 2 SDs of the mean (here the unknown population proportion), so this method will “capture” the process probability for roughly 95% of samples.

However, we should admit that the multiplier of 2 is a bit of a simplification. So how do we find a more precise value of the multiplier to use, including for values other than 95%? We will use technology to do this. Keep in mind that the  $z$ -value corresponding to probability  $C$  in the middle of the distribution, also corresponds to having probability  $(1 - C)/2$  in each tail.

### Technology Detour – Finding Percentiles from the Standard Normal Distribution

#### Normal Probability Calculator applet

- You can leave the mean set to zero and the standard deviation to 1 and the variable is “z-scores.”
- Check the box next to the first  $<$  sign and specify the *probability value* to correspond to the lower tail probability of interest,  $(1 - C)/2$ . Press Enter/Return and it will display the negative  $z$ -value.

In R: The `iscaminvnorm` function takes the following inputs:

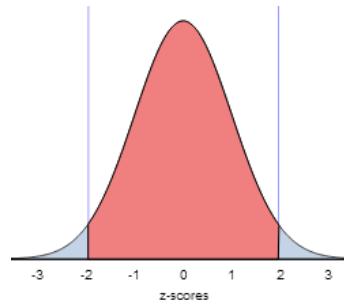
- *Probability* of interest
- *Mean* = the mean of the normal distribution (default = 0)
- *Standard deviation* = the standard deviation of the normal distribution (default = 1)
- *Direction* = whether the probability of interest was in the lower tail (“below”) or the upper tail (“above”), in both tails (“outside”), or in the middle of the distribution (“between”)

For example: `iscaminvnorm(prob=.95, direction = "between")`

- (f) Use technology to more precisely determine the number of standard deviations that capture the middle 95% of the normal distribution with mean = 0 and standard deviation = 1. [Hint: In other words, how many standard deviations do you need to go on each side of zero to capture the middle 95% of the distribution?]

**Definition:** The  $(100 \times C)\%$  [critical value](#),  $z^*$ , is the  $z$ -score value such that  $P(-z^* \leq Z \leq z^*) = C$  where  $C$  is any specified probability value, and  $Z$  represents a normal distribution with mean 0 and SD 1.

Note: We use the symbol  $z^*$  to distinguish this value, found based on the confidence level, from  $z_0$ , the observed  $z$ -score for the data.



- (g) Find the critical value for a 90% confidence interval. Is it larger or smaller than with 95% confidence? Why does this make sense?

Putting these together, we have a [one-sample  \$z\$ -interval](#) (aka Wald interval) for a process probability:

**One sample  $z$ -confidence interval (or “Wald interval”) for  $\pi$ :** When we have at least 10 successes and at least 10 failures in the sample, an approximate confidence interval for  $\pi$  is given by:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- (h) Based on this formula, what is midpoint of the interval? What determines the width of the interval?

Midpoint:

Width:

- (i) How does increasing the confidence level affect the width of the interval?

How does increasing the sample size affect the width of the interval?

- (j) Use this procedure to calculate (by hand) [and interpret](#) a 90% confidence interval, based on the normal probability model, for the probability a kissing couple leans to the right.

- (k) How do the midpoint and width of the 90% confidence interval compare to those of the 95% confidence interval? Also verify your claims in (h) about the midpoint and width based on this 90% confidence interval.

**Definition:** The *half-width* of the interval is also referred to as the [margin-of-error](#).

So the above interval formula is of the common form:  $statistic \pm margin-of-error$ , where  $margin-of-error = critical\ value \times standard\ error\ of\ statistic$ .

Although 95% is the most common confidence level, a few other confidence levels and their corresponding critical values are shown in the table below.

Confidence level	90%	95%	99%	99.9%
Critical value $z^*$	1.645	1.960	2.576	3.291

### Technology Detour – One proportion z-confidence intervals

#### z-interval with R

- Use `iscamonepropztest(observed, n, hypothesized, alternative="greater", "less," or "two.sided", conf.level)`

You can enter either the number of successes or the proportion of successes ( $\hat{p}$ ) for the “observed” value. If you don’t specify a hypothesized value and alternative, be sure to label the confidence level

For example: `iscamonepropztest(observed=80, n=124, conf.level=95)`

#### z-interval with [Theory-Based Inference](#) applet

- Keep the pull-down menu set to **One proportion**.
- Specify the sample size  $n$  and either the count or the sample proportion. Or check the Paste Data box and paste in the individual outcomes (check the **Includes header** box if you are also copying over the variable name).
- Press **Calculate**.
- Check the box for **Confidence interval**
- Change the **confidence level** from 95 to 90 %
- Press the **Calculate CI** button

Scenario: One proportion

Paste Data

n: 124  
count: 80  
sample  $\hat{p}$ : 0.6452

Confidence interval

confidence level 90 %

(l) Now use technology to find and report the 90% and 95% confidence intervals for the probability that a kissing couple leans to the right. Did the widths and midpoints behave as you conjectured above?

(m) How do these intervals compare to the Exact Binomial Confidence Intervals reported by R?

90 percent confidence interval: 0.5683679 0.7166460

95 percent confidence interval: 0.5542296 0.7289832

## Sample Size Determination

(n) Suppose you are planning your own study about kissing couples. Before you collect the data, you know you would like the margin-of-error to be at most 3 percentage points and that you will use a 95% confidence level. Use this information to determine the sample size necessary for your study.

[This is a very common question asked of statisticians. Think about how to determine this using the z-interval formula and whether you would answer differently based on whether you have any prior guess about  $\pi$ . In this case, use the sample proportion found in the original study as an estimate for the unknown value of  $\pi$ . Without a preliminary study, you can use 0.5 as an estimate of this probability, in order to produce a conservative estimate that makes the required sample size as large as possible. Compare these results. Also think about how much more difficult this question would be to answer using the binomial distribution.]

Using sample proportion:

Using 0.5 for  $\pi$ :

## Interpretation of Confidence Level

The next few questions will explore further what is meant by the phrase “95% confidence.” To do so, we will take a hypothetical situation where we know the value of the process probability and then use simulation to see how the procedure does at estimating this value  $\pi$ . This will aid our attempt to answer the more realistic question of “does my interval actually contain the value of the unknown parameter?”

(o) Open the [Simulating Confidence Intervals](#) applet. Assume that  $\pi = 0.667$  is the probability that a kissing couple leans to the right.

- Keep the first pull-down menu set to **Proportions**.
- Keep the second pull-down menu set to **Binomial** and the third set to **Wald** (another name for the one sample z-interval).
- Specify 0.667 as the process probability  $\pi$  and 124 as the sample size **n**.
- Keep the confidence level (**Conf level**) set to 95%.
- Press **Sample** to produce a randomly generated sample proportion from a process with  $\pi = 0.667$  and its corresponding Wald confidence interval for  $\pi$ .

**Method:**

Proportions
Binomial
Wald
$\pi$ 0.667
n 124
Intervals 1
<b>Sample</b>

---

Conf level 95 %

Did this interval capture the value of  $\pi$  (0.667)? [Hint: The interval will be green if 0.667 is inside the interval and red if not.]

(p) Click on the interval itself, and the midpoint of the interval will display (at the bottom), as well as the endpoints. Record these values.

Midpoint:

Lower endpoint:

Upper endpoint:

(q) Press **Sample** again. Did you get the same value of  $\hat{p}$ ?

New interval:

Did you get the same interval?

Does this surprise you?

Did the value of  $\pi$  change?

Did the new interval capture the value of  $\pi$ ?

(r) Now change the number of intervals (**Intervals**) to **198** (for a total of 200 intervals) and press **Sample**. What percentage of these 200 random intervals successfully capture 0.667 (see the **Running Total**)?

Is this close to what you expected?

(s) Also press **Sort**, and comment on what the intervals that fail to successfully capture 0.667 have in common.

(t) Change the number of intervals to **200** and press **Sample** four more times. Examine the Running Total for these 1000 intervals. Is this percentage (of intervals that successfully contain the value underlying parameter value 0.667) close to what you expect?

**Definition:** The [confidence level](#) of an interval procedure is supposed to indicate the long-run percentage of intervals that capture the actual parameter value, if repeated random samples were taken from the process. As such the confidence level presents a measure of the *reliability* of the method. A confidence interval procedure is considered valid when the achieved long-run coverage rate matches (or comes close to) the stated (also called nominal) confidence level.

- (u) Based on your simulation results above, does this one-sample  $z$ -interval appear to be a valid 95% confidence interval procedure in this case? Explain.
- (v) Predict what will be different about 90% confidence intervals compared to the 95% confidence intervals. (*Hint:* Try to list two differences.)
- (w) Press the **Sort** button and then change the **confidence level** to **90%** and press **Recalculate**. What is the primary change in the properties of these intervals?
- In particular, does the coverage rate appear to now differ from 95%? Explain.
- (x) Predict what will be different about these 90% confidence intervals if we double the sample size. Also predict what (if anything) will be different about the running percentage of intervals that succeed in capturing the parameter value.
- (y) Change the sample size **n** to **248** (twice as large as before) and press **Sample** five times to generate 1000 intervals. What is the primary change in the properties of these intervals?

Does the (long-run) confidence level appear to now differ from 90%?

**Discussion:** In real life, you only take one sample, and you cannot know whether it's going to produce a green or a red interval! So what does "95% confidence" mean? These simulations reveal that if we take repeated random samples from a process with probability  $\pi$  and calculate a  $(100 \times C)\%$  confidence interval each time, then in the long-run roughly  $(100 \times C)\%$  of the resulting intervals succeed in capturing  $\pi$  inside the interval. So when we gather the actual sample data and calculate a *single* interval, we say that we are " $(100 \times C)\%$  confident" that this interval succeeds in capturing the actual value of the process probability  $\pi$ ; and the basis for our confidence is knowing that in the long run the procedure would generate a successful interval in  $(100 \times C)\%$  of all possible samples. We also noted that sample size will affect the width or "precision" of the interval but not the confidence level.

- (z) Suppose that you take 1000 random samples from a process and generate a 95% confidence interval for  $\pi$  with each sample. If you randomly select one of these intervals, what is the probability that you will select an interval that captures  $\pi$ ?

- (aa) Why is it technically incorrect to say “If I have calculated a 95% confidence interval for  $\pi$  to be (0.641, 0.738), there is a 0.95 probability that  $\pi$  is between 0.641 and 0.738.”?

**Practice Problem 1.10**

- (a) Determine and interpret the  $z$ -confidence interval for Muffin from Practice Problem 1.9A (treat the blue ball as success). Is this procedure valid in this situation? How are you deciding?
- (b) What would be the necessary sample size if we wanted a margin-of-error of 0.01 for a confidence level of 95%? Explain how you are finding this.
- (c) In an actual study, how do you know whether your interval actually contains the value of the unknown parameter (that is, whether it is a red or a green interval)?
- (d) What is the distinction between *standard deviation* and *standard error*?

### Investigation 1.11: Heart Transplant Mortality (cont.)

Recall Investigation 1.3, where you learned that 8 of the 10 most recent heart transplantation operations at St. George's Hospital resulted in a death.

- (a) Use the one sample  $z$ -interval method to find a 95% confidence interval for the probability of a heart transplantation death at St. George's hospital. Does anything bother you about doing this?

To consider whether the above procedure is valid, we can again pretend we know the process probability, generate many random samples and calculate confidence intervals from them, and see what percentage of these confidence intervals capture the actual value.

- (b) In the [Simulating Confidence Intervals](#) applet, suppose the actual process probability of death is 0.15 and you plan to take a sample of 10 operations and apply the  $z$ -interval procedure. Use the applet to explore the reliability (empirical coverage rate) of this method for these data. That is, generate 1000 intervals (200 at a time) with  $\pi = 0.15$  and  $n = 10$  and see how many of these intervals succeed in capturing the actual value of the population parameter (0.15). Is this coverage rate close to 95%?

- (c) Explain why you should not expect the coverage rate to be close to 95% in this case.

Because the  $z$ -interval procedure has the sample size conditions of the Central Limit Theorem, there are scenarios where we should not use this  $z$ -procedure to determine the confidence interval. One option is to return to calculating a confidence interval based on the Binomial distribution. However, this is a fairly complicated method (as opposed to: “go two standard deviations on each side”) and often produces intervals that are wider than they really need to be (the actual coverage rate is higher than the nominal confidence level). An alternative method that has been receiving much attention of late is often called the [Plus Four](#) procedure (in contrast to the *Wald* procedure we’ve been using): The idea is to pretend you have 2 more successes and 2 more failures than you really did (the “Wilson adjustment”).

**Definition: Plus Four 95% confidence interval for  $\pi$ :**

- Determine the number of successes ( $X$ ) and sample size ( $n$ ) in the study
- Increase the number of successes by two and the sample size by four. Make this value the midpoint of the interval:  $\tilde{p} = (X + 2)/(n + 4)$
- Use the  $z$ -interval procedure as above for the augmented sample size of  $(n + 4)$ :

$$\tilde{p} \pm 1.96 \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

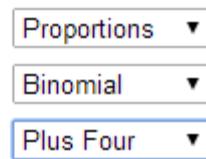
*Note:* For confidence levels other than 95%, researchers have recommended using the *Adjusted Wald* method:  $\tilde{p} = (X + 0.5z^*{}^2)/(n + z^*{}^2)$  and  $\tilde{n} = n + z^*{}^2$  so the interval becomes  $\tilde{p} \pm z^* \sqrt{\tilde{p}(1 - \tilde{p})/\tilde{n}}$ .

(d) Investigate the reliability of the Plus Four procedure:

In the **Simulating Confidence Intervals** applet

- Use the second pull-down menu to change from Wald to **Plus Four**.
- Generate 1000 confidence intervals from a process with  $\pi = 0.15$  and  $n = 10$ .

Method:



Is this coverage rate close to 95%? Is this an improvement over the (Wald)  $z$ -interval?

(e) Use the second pull-down menu to toggle between the Wald and Plus Four intervals. [Hint: You can **Sort** the intervals first.] What do you notice about how the *midpoints* of the intervals compare? (Also think how  $\tilde{p}$  and  $\hat{p}$  differ.) Why is that useful in this scenario?

Researchers have determined that this method does have much better coverage properties, even for very small sample sizes, and tends to produce intervals that are narrower than the binomial intervals. Some statisticians have gone so far as to argue that this is the only procedure that should be used for a confidence interval for a process probability.

(f) Use the Plus Four procedure to determine and interpret a 95% confidence interval for the probability of a death during a heart transplant operation at St. George's hospital. [Hints: You can do the calculation either by hand, first finding  $\tilde{p}$  and  $z^*$ , or with the Theory-Based Inference applet or Minitab or R by telling the technology there were 4 more operations consisting of two more deaths than in the actual sample.]

(g) Repeat for the larger study at St. George's that revealed 71 deaths out of 361 operations. How do the 95% Wald and Plus Four intervals compare in this case?

Wald interval:

Plus Four interval:

Comparison:

**Discussion** As you can tell, there are several ways to obtain a confidence interval for a process probability. When the sample size is large, they will yield very similar results for the endpoints and the coverage rate. When the sample size is small, the “Adjusted Wald” method is preferred (or the Binomial but it tends to be more conservative and doesn’t have the *estimate ± margin-of-error* simplicity). Below we compare the 95% confidence intervals for the 8 of 10 and 71 of 361 studies for these methods plus a fourth, the “Wilson interval” which is a more direct inversion of the one sample proportion z-test.

	<u><math>n = 10</math></u>	<u><math>n = 361</math></u>
• <b>Exact Binomial</b> (aka Clopper-Pearson) Finds the values of $\pi$ so the $P(X \leq k) \geq (1 - C)/2$ and $P(X \geq k) \geq (1 - C)/2$ where $k$ is observed number of successes Tends to be “conservative” (longer than necessary)	(.4439, .9748)	(.1569, .2415) Minitab or R using iscaminomtest
• <b>Wald interval</b> (aka normal approximation aka asymptotic) Finds the values of $\pi$ so that $P(-z \leq \frac{\hat{p} - \pi}{\sqrt{\hat{p}(1 - \hat{p})/n}} \leq z) = C$ Formula: $\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})/n}$ Tends to show poor coverage properties if small $n$ or extreme $\pi$	(.5521, 1.048)	(.1557, .2377) TBI applet or Minitab or R (iscamonepropztest)
• <b>Plus Four</b> (aka Agresti-Coull, special case of Adjusted Wald) (95%) Add two successes and two failures to sample results Formula: $\tilde{p} \pm z^* \sqrt{\tilde{p}(1 - \tilde{p})/\tilde{n}}$ where $\tilde{p} = (X + 0.5z^2)/(n + z^2)$ and $\tilde{n} = n + z^2$ Very good coverage properties	(.4776, .9509)	(.1590, .2410) TBI applet or Minitab or R after making adjustment to sample data
• <b>Wilson interval</b> (aka Score interval) Finds the values of $\pi$ so that $P(-z \leq \frac{\hat{p} - \pi}{\sqrt{\pi(1 - \pi)/n}} \leq z) = C$ Formula: $\frac{\hat{p} + \frac{1}{2n}z^2 \pm z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + z^2/n}$ Very good coverage properties. Very similar to Adjusted Wald for 95% confidence.	(.4902, .9433) R prop.test w/o continuity corr.	(.1590, .2408) (.4422, .9646) R prop.test w/ continuity corr. (.1577, .2423)

Many statisticians now view the Plus Four or Adjusted Wald method as the best in terms of coverage rate and width (Binomial intervals, because of the discreteness, are often wider than they “need” to be), as well as simplicity.

Whatever procedure is used to determine the confidence interval, you interpret a (valid) interval the same way – as the interval of plausible values for the parameter. For example, we are 95% confident that the underlying probability of death within 30 days of a heart transplant operation at St. George’s Hospital is between 0.16 and 0.24; where by “95% confident,” we mean that if we were to use this

procedure to construct intervals from thousands of representative samples, roughly 95% of those intervals will succeed in capturing the actual (but unknown) value of the parameter of interest.

**Keep in Mind:** The main things you should focus on in your study of confidence intervals are:

- What parameter is the interval estimating (in context)?
- What are the effects of sample size, confidence level, and  $\hat{p}$  on the width and midpoint interval?
- What do we mean by “confidence”?
- What are the effects (if any!) of sample size, confidence level, and  $\hat{p}$  on the coverage rate of the method?
- Why might one confidence interval method be preferred over another?

Also note how to use R, Minitab, or the Theory-Based Inference applet to perform these calculations. You should NOT use the Simulating Confidence Intervals applet to construct a confidence interval for a particular sample of data.

### Practice Problem 1.11

Return to your exploration using the [Simulating Confidence Intervals](#) applet. Assume that  $\pi = 0.667$  is the probability that a kissing couple leans to the right.

(a) From the third pull-down menu select the Wilson interval method. Evaluate the performance of this interval method, clearly explaining your steps.

(b) Starting with  $z = \frac{\hat{p} - \pi}{\sqrt{\pi(1-\pi)/n}}$ , use the quadratic formula to solve for  $\pi$  and verify the formula for the Score interval.

(c) Show that with 95% confidence the Wilson formula simplifies to approximately the Plus Four method.

### SECTION 3: SAMPLING FROM A FINITE POPULATION

So far, we have treated each set of data that we have investigated as a sample from an ongoing process with an underlying process probability. It is more common to consider sample data as coming from a larger, but finite, population (e.g., a sample of adults from among all Americans aged 18 years and older). So what we now want to consider are (i) how to select a sample from the population to allow us to generalize our sample observations back to that larger population and (ii) what statistical techniques we need to make such inferences. You will see that though many of the analysis procedures are the same, there are a few more issues to consider as well.

#### Investigation 1.12: Sampling Words

- (a) Circle 10 representative words in the following passage.

*Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal.*

*Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.*

*We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.*

*But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead, who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here.*

*It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.*

The authorship of literary works is often a topic for debate. Were some of the works attributed to Shakespeare actually written by Bacon or Marlowe? Which of the anonymously published *Federalist Papers* were written by Hamilton, which by Madison, which by Jay? Who were the authors of the writings contained in the Bible? The field of “literary computing” examines ways of numerically analyzing authors’ works, looking at variables such as sentence length and rates of occurrence of specific words.

The above passage is of course Abraham Lincoln’s Gettysburg Address, given November 19, 1863 on the battlefield near Gettysburg, PA. In characterizing this passage, we would ideally examine every word. However, often it is much more convenient and even more efficient to only examine a subset of words.

**Definition:** The population is the entire collection of observational units that we are interested in. A sample is the subset of the population for which we collect data.

In this case you will examine data for just 10 of the words. We are considering this passage to be a *population* of 268 words, and the 10 words you selected are therefore a *sample* from this population.

In most statistical studies, we do not have access to the entire population and can consider only data for a sample from that population. Our ultimate goal is to make conclusions about the larger population, based on only the sample data.

Up until now, we have generally been assuming we have a representative sample from an infinite on-going process (e.g., dog identification of cancer sample, hospital transplant operations, candy manufacturing). We made some assumptions, like the process is not changing over time and that there is no tendency to select some types of outcomes more than others (e.g., getting the first 5 candies from the manufacturing process rather than throughout the day). In fact, a binomial process assumes you have repeat observations from the exact same process but with randomness in the actual outcome that occurs (sometimes the dog makes the correct identification, some time she does not).

In this case, instead of sampling from an on-going process we are sampling from a finite population (the 268 words). In fact, we actually have access to the entire population. But what if we didn’t? We still need some way of convincing people that a sample is likely to be representative of the population. To do that, we will explore for a minute, using this population to see how samples behave where “random chance” arises from which observational units are selected to be in the sample, rather than from “random choices” made by the observational units.

(b) Consider the sample you selected in (a). Suppose I wanted to focus on *the lengths of the words* in the sample (if the sample is truly representative, it shouldn’t matter what variable I end up recording). Record the ten lengths that you found.

	1	2	3	4	5	6	7	8	9	10
Length										

(c) Identify the observational units and variable for this sample.

Obs units:

Variable:

Type:

- (d) Do you think the words you selected are representative of the population of 268 words in this passage? How are you deciding?

**Definition:** The term [parameter](#), before considered the process probability, is also used to refer to a numerical characteristic of a *population*. We will continue to denote population parameters with Greek letters, for example  $\pi$  or  $\mu$  for a population proportion or population mean, respectively. A statistic continues to be the corresponding number but calculated from sample data. We denote the statistics for a sample proportion and a sample mean by  $\hat{p}$  and  $\bar{x}$ , respectively.

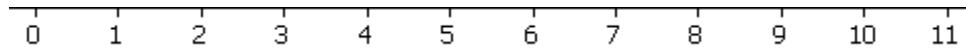
- (e) Calculate the average length of the ten words in your sample. Is this number a parameter or a statistic? What symbol can we use to refer to it?

Average: \_\_\_\_\_ Parameter or statistic? Symbol:

- (f) The average length of all 268 words in this population is 4.29 letters. Is this number a parameter or a statistic? What symbol can we use to refer to it?

- (g) Is your sample average similar to the population average? Did everyone in your class obtain the same sample average? Describe a way for deciding whether this sampling method tends to produce samples that are representative of the larger population.

- (h) Construct a dotplot displaying the *average length of words* in your sample and those of your classmates.



Label:

What are the observational units in this graph?

- (i) What does the graph in (h) tell you about whether this sampling method tends to produce representative samples?

**Discussion:** You have again witnessed the fundamental principle of [sampling variability](#): Values of sample statistics *vary* when one takes different samples from the same population. A key point in analyzing these results is that now we are treating the samples as the observational units and the sample statistics as the variable of interest (so the graph label should be something like “sample mean word lengths”). Although we clearly expect there to be sample-to-sample variation in the statistic, the problem is if there is a tendency to over or underestimate the parameter of interest.

**Definition:** When characteristics of the resulting samples are systematically different from characteristics of the population, we say that the sampling method is [biased](#). When the distribution of the sample statistics, under repeated sampling from the same population, is centered at the value of the population parameter, the sampling method is said to be [unbiased](#).

For example, we suspect that your class repeatedly and consistently overestimated the average length of a word. Not everyone has to overestimate, but if there is a *tendency* to err in the same direction time and time again, then the sampling method is biased. In other words, sampling bias is evident if we repeatedly draw samples from the population and the distribution of the sample statistics is not centered at the population parameter of interest. Note that bias is a property of a sampling *method*, not of a single sample. Studies have shown that human judgment or “[convenience sampling](#)” (e.g, selecting the most readily available observational units) is not a good basis for selecting representative samples, so we will rely on other techniques to do the sampling for us.

(j) Does your class results indicate a tendency to overestimate or to underestimate the population mean  $\mu$ ? Could you have predicted that in advance? Explain.

(k) Consider another sampling method: you close your eyes, point at the passage of words, select whatever word your pen lands on, and repeat this 10 times. Would this sampling method be biased? If so, in which direction? Explain. What if you looked at 20 words instead of 10?

(l) Suggest a better method for selecting a sample of 10 words from this population that is likely to be representative of the population.

**Definition:** A [simple random sample](#) gives every observational unit in the population the same chance of being selected. In fact, it gives every sample of size  $n$  the same chance of being selected.

So, with a simple random sample, any set of 10 words is equally likely to end up as our sample.

“Low-tech” methods for obtaining a simple random sample from a population include using a *random digits table*. A random digits table is constructed so that each position is equally likely to be filled by any digit from 0 to 9, and the digit in one position is unaffected by the digit in any of the other positions.

The first step is to obtain a list of every member of your population (this list is called a [sampling frame](#)). Then, give each observational unit on the list a unique ID number.

The following is a sampling frame for the Gettysburg address, with each word in the population numbered. You then use a random digits table or technology to select a 3-digit number at random, and then match those to the words in the population with those IDs.

1 Four	35 in	69 dedicate	103 But,	137 add	171 here	205 these	239 that
2 score	36 a	70 a	104 in	138 or	172 to	206 honored	240 this
3 and	37 great	71 portion	105 a	139 detract.	173 the	207 dead	241 nation,
4 seven	38 civil	72 of	106 larger	140 The	174 unfinished	208 we	242 under
5 years	39 war,	73 that	107 sense,	141 world	175 work	209 take	243 God,
6 ago,	40 testing	74 field	108 we	142 will	176 which	210 increased	244 shall
7 our	41 whether	75 as	109 cannot	143 little	177 they	211 devotion	245 have
8 fathers	42 that	76 a	110 dedicate,	144 note,	178 who	212 to	246 a
9 brought	43 nation,	77 final	111 we	145 nor	179 fought	213 that	247 new
10 forth	44 or	78 resting	112 cannot	146 long	180 here	214 cause	248 birth
11 upon	45 any	79 place	113 consecrate,	147 remember,	181 have	215 for	249 of
12 this	46 nation	80 for	114 we	148 what	182 thus	216 which	250 freedom,
13 continent	47 so	81 those	115 cannot	149 we	183 far	217 they	251 and
14 a	48 conceived	82 who	116 hallow	150 say	184 so	218 gave	252 that
15 new	49 and	83 here	117 this	151 here,	185 nobly	219 the	253 government
16 nation:	50 so	84 gave	118 ground.	152 but	186 advanced.	220 last	254 of
17 conceived	51 dedicated,	85 their	119 The	153 it	187 it	221 full	255 the
18 in	52 can	86 lives	120 brave	154 can	188 is	222 measure	256 people,
19 liberty,	53 long	87 that	121 men,	155 never	189 rather	223 of	257 by
20 and	54 endure.	88 that	122 living	156 forget	190 for	224 devotion,	258 the
21 dedicated	55 We	89 nation	123 and	157 what	191 us	225 that	259 people,
22 to	56 are	90 might	124 dead,	158 they	192 to	226 we	260 for
23 the	57 met	91 live.	125 who	159 did	193 be	227 here	261 the
24 proposition	58 on	92 it	126 struggled	160 here.	194 here	228 highly	262 people,
25 that	59 a	93 is	127 here	161 it	195 dedicated	229 resolve	263 shall
26 all	60 great	94 altogether	128 have	162 is	196 to	230 that	264 not
27 men	61 battlefield	95 fitting	129 consecrated	163 for	197 the	231 these	265 perish
28 are	62 of	96 and	130 it,	164 us	198 great	232 dead	266 from
29 created	63 that	97 proper	131 far	165 the	199 task	233 shall	267 the
30 equal.	64 war.	98 that	132 above	166 living,	200 remaining	234 not	268 earth.
31 Now	65 We	99 we	133 our	167 rather,	201 before	235 have	
32 we	66 have	100 should	134 poor	168 to	202 us,	236 died	
33 are	67 come	101 do	135 power	169 be	203 that	237 in	
34 engaged	68 to	102 this.	136 to	170 dedicated	204 from	238 vain,	

## Technology Detour – Selecting a Simple Random Sample

**In applet** You may can [random.org](#) or our [Generate Random Numbers](#) applet to select a simple random sample of integers by specifying the range of integers (e.g., 1 to 268) and, if an option, how many numbers you want to generate.

### In R

- Create a vector of integers from 1 to 268

In the Console window:

```
> ids = 1:268
```

- Take a simple random sample of 5 ID numbers:

```
> sample(ids, 5)
```

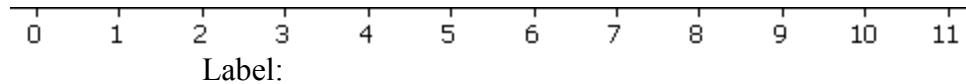
[Use the up arrow to repeat this command if you are sharing computers. If you are using the ISCAM workspace, you should also reset the random number seed: `rm (.Random.seed)` ]

(m) Select a simple random sample of 5 ID numbers from the sampling frame. Match these ID numbers with the corresponding words. Write down the resulting ID numbers, the selected words, and the length of each word.

	1	2	3	4	5
ID number					
Word					
Length					

(n) Calculate the average word length in your random sample of 5 words.

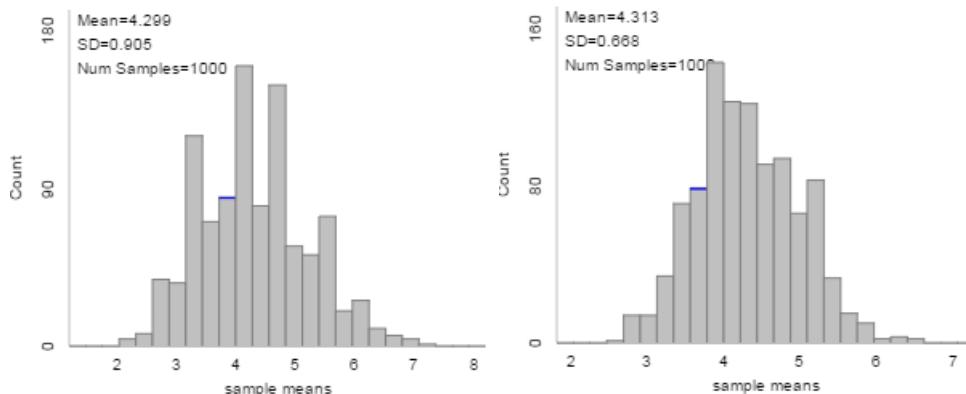
(o) Again combine your results with your classmates to produce a dotplot of the sample mean word lengths. Comment on the distribution and particularly how it compares to the one in (h).



(p) When taking random samples, did everyone obtain a sample mean equal to the population mean or is there still sample to sample variation?

(q) So then in what way would you say these random samples produce “better” sample results than the “circle 10 words” method?

Of course, to really see the long-term pattern to the sampling distribution of the statistic, we would like to take many more random samples from this population. The following dotplots display the sample means from 1000 random samples from this population.



(r) For one graph, there were 5 words in each sample; for the other graph there were 10 words. Which is which? How are you deciding?

(s) Would you consider both of these graphs to have arisen from unbiased sampling methods?

**Discussion:** Even with randomly drawn samples, sampling variability still exists (*random sampling errors*). Although not every random sample produces the same characteristics as the population, random sampling has the property that the sample statistics do not consistently overestimate the value of the parameter or consistently underestimate the value of the parameter (which is not changing). The means of the above graphs of sample means (the statistic) are both very close to 4.29, the population mean (the parameter). Because the distributions of the sample statistic (as a random variable) are centered at the value of the population parameter, both graphs are considered to demonstrate *unbiased* sampling methods. In fact, one can show mathematically that random sampling is always an unbiased sampling method with a sample mean:  $E(\bar{X}) = \mu$  (but may not be unbiased for other choices of statistics!). Notice that in judging bias we are concerned only with the mean of the distribution, not the shape or variability. However, if we do take larger samples ( $n = 10$  rather than  $n = 5$ ), we do improve the *precision* of our statistics as they will tend to cluster even more tightly around the population mean. In designing a study, it is critical to first make sure there is no sampling bias (use random sampling), then the sample size (not the number of samples) will determine how close the statistic can be expected to be to the parameter.

## Summary

We have seen two very desirable properties of sample means when we use *random samples* from a large population:

- The distribution of sample means centers at the population mean (so the sampling method is *unbiased*).
- The distribution of sample means has less sample to sample variability when we increase the sample size (producing increased *precision*).

**Practice Problem 1.12A**

Suppose that you take many random samples (i.e., thousands) from a population and graph the distribution of the resulting sample statistics.

- (a) If the distribution of sample statistics is centered at the value of the population parameter, then is the sampling method unbiased? (*Choose one:* Yes definitely, Maybe or maybe not, No definitely not)
- (b) If the distribution of sample statistics appears to be normally distributed, then is the sampling method unbiased? (*Choose one:* Yes definitely, Maybe or maybe not, No definitely not)
- (c) If most of the sample statistics are close to the value of the population parameter, then is the sampling method unbiased? (*Choose one:* Yes definitely, Maybe or maybe not, No definitely not)
- (d) If the sampling method is biased, then will increasing the sample size reduce the bias? (*Choose one:* Yes definitely, Maybe or maybe not, No definitely not)

**Practice Problem 1.12B**

In this investigation, we found that the *average* of the sample *average* is near the population *average*. Explain what each use of the term “average” means in this statement.

**Practice Problem 1.12C**

Explain the distinctions between the following pairs of terms:

- (a) parameter and statistic
- (b) bias and precision
- (c) sample size and number of samples
- (d) sampling error and random sampling error

**Investigation 1.13: *Literary Digest***

*Literary Digest* was a well-respected political magazine founded in 1890. Using sampling, they correctly predicted the presidential outcomes from 1916–1932. In 1936, they conducted the most extensive (to that date) public opinion poll in history. They mailed out questionnaires to over 10 million people (about one-third of U.S. households) whose names and addresses they obtained from telephone books and vehicle registration lists. More than 2.4 million responded, with 57% indicating that they would vote for Republican Alf Landon in the upcoming presidential election.

- (a) Identify the population of interest, the sample, and the sampling frame in this study. Also define the parameter and statistic, in words and symbols, and indicate any values that you know.

Population of interest

Sample

Sampling frame

Parameter

Statistic

- (b) Have you ever heard of Alf Landon? He lost. By a landslide. Incumbent Democrat Franklin Roosevelt won the election, carrying 63% of the popular vote to Landon's 37%. Give two explanations why the *Literary Digest* prediction was so much in error. In particular, talk about the direction of the bias – why was this sampling method vulnerable to producing an overestimate of the parameter?

**Discussion:** There are two main issues here. One, the sampling frame did not include all members of the population of interest and failed to include those that were poorer (and at that time likely to be Democrat). Second, the voluntary response nature of the poll implies the surveyors were more likely to hear from those unhappy with the status quo or with more time on their hands to complete such surveys (e.g., retired folks), and even those more willing to pay for a stamp. All of these probably point to an overrepresentation of Republicans. Bad sampling frames and voluntary response bias are perhaps the most common sources of sampling error. By the way, a fledgling pollster of the time, George Gallup, actually bet that he would predict the percentages more accurately. Not only did he correctly predict the *Digest* result with only 3,000 respondents, he also correctly predicted a Roosevelt victory!

**Practice Problem 1.13A**

In the mid-1980s, Shere Hite undertook a survey of women's attitudes toward relationships, love, and sex by distributing 100,000 questionnaires in women's groups. Of the 4500 who returned the questionnaire, 96% said that they gave more emotional support than they received from their husbands or boyfriends. An ABC News/Washington Post poll surveyed a random sample of 767 women, finding that 44% claimed to give more emotional support than they received.

(a) Which of the following is the “parameter” of interest in these studies?

- Those who said they gave more emotional support
- All American women
- The 4500 women who responded
- The proportion who said they received less emotional support
- Other

(b) Which poll’s results do you think are more representative of the population of all American women? Explain.

**Practice Problem 1.13B**

An article published in the June 6, 2006 issue of the journal *Pediatrics* describes the results of a survey on the topic of college students intentionally injuring themselves. Researchers invited 8300 undergraduate and graduate students at Cornell University and Princeton University to participate in the survey. A total of 2875 students responded, with 17% of them saying that they have purposefully injured themselves. Suppose we are interested in the proportion of self-injuries in the population of all college students.

- (a) Identify the observational units and variable in this study. Also classify the variable as categorical (also binary?) or quantitative.
- (b) Identify the parameter and statistic. Also indicate appropriate symbols.
- (c) Do you think it likely that this sample is representative of the population of all college students in the world? What about all college students in the U.S.? Explain.
- (d) Describe a population you might be willing to generalize the results to.
- (e) For which of the following variables would you suspect this sample would be representative of the population of all U.S. college students? Justify your answer.

Favorite TV show

Height

Parental education

Eye color

**Practice Problem 1.13C**

- (a) Does your class constitute a random sample of students from your school? Explain why or why not.
- (b) Suggest a variable for which your class should *not* be considered a representative sample of all students at your school. Explain why not.
- (c) Suggest a variable for which it might be reasonable to consider your class to be representative of all students at your school. Justify your choice.

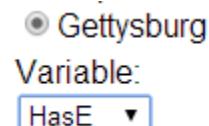
### Investigation 1.14: Sampling Words (cont.)

In this section, we have transitioned from sampling from an infinite process to sampling from a finite population. We discussed *randomly selecting* the sample from a list of the entire population as a way to convince us that the sample is likely to be representative of the larger population and therefore we are willing to *generalize* results from our sample to the larger population. The second benefit of consciously introducing this randomness into the study design is that the resulting sample statistics will follow a predictable pattern as suggested in Investigation 1.12, allowing us to measure the strength of evidence against claims about the population parameter and the margin-of-error in our estimates of the parameter. We will again consider simulation-based p-values, exact p-values, and a mathematical model for estimating the p-value and confidence interval.

#### Statistical Inference

To calculate p-values and confidence intervals, we would like to examine all possible samples of size  $n$  from the population and calculate the statistic for each sample to understand the amount of sample-to-sample variation by chance alone. However, this is not usually feasible. Even for samples of size 5 from a population of 268 words, there are  $C(268, 5)$  or more than 11 billion possible samples! Instead, we will use technology to take a large number of random samples to estimate the behavior of the distribution of the sample statistics in this rare case where we have access to the entire population.

- (a) Open the [Sampling Words](#) applet. You will notice that the applet displays a histogram of word lengths for the entire population. Use the **Variable** pull-down menu to select **HasE** to focus on the variable of whether or not the word contains at least one *e*.



The applet displays the distribution of “*e*-words” and “non-*e*-words.” Is this variable quantitative or categorical?

- (b) What is the value of the parameter (proportion of *e*-words) for this population? What symbol can we use to represent this value?

- (c) Check the **Show Sampling Options** box. Specify 1000 as the **Number of Samples** and 5 as the **Sample Size**. Press the **Draw Samples** button. Describe the resulting distribution of sample proportions. Be sure to explain what the mean and standard deviation represent here.

- (d) Does the random sampling performed by this applet appear to be an *unbiased* sampling method to estimate the population proportion? Explain how you are deciding.

## Effect of Sample Size

(e) Change the sample size to 20 press Draw Samples. Discuss how this change has impacted the bias of the sampling method and/or the precision. [Hint: Examine the reported mean and standard deviation.] Is this what you expected?

## Effect of Population Size

(f) Now select the **x4** radio button on the left to make four copies of the population Your population now consists of the four copies of the Gettysburg Address, a total of 1072 words. Now what is the value of the population proportion of successes ( $e$ -words)?

(g) Generate 1000 random samples of  $n = 20$  words from this new population, and observe the distribution of 1000 sample proportions. Compare how the shape, mean, and standard deviation compare to that from (e).

Mean:

Standard deviation:

Shape:

Comparison:

Is this what you expected?

(h) Repeat (g) using **forty addresses** as the population, now containing 10,720 words.

Mean:

Standard deviation:

Shape:

**Discussion:** You should notice that the population size did not make much of a difference in the distribution of sample proportions! This will be true as long as the population size is large compared to the sample size. In this case, the small sample we are removing from the population is not changing the population much (so the probability the next word contains the letter  $e$  is pretty much the same as the probability the first word selected would) and we can consider the observations to be approximately independent (the subsequent probability of success to be approximately constant). This means it's just like we are sampling from an infinite process...

(i) In Investigation 1.7, we stated that the mean of a distribution of sample proportions from a process equals  $\pi$  and the standard deviation equals  $\sqrt{\pi(1-\pi)/n}$ . Calculate these theoretical values and compare them to the applet results for  $n = 20$ .

Theoretical Mean:

Theoretical Standard deviation:

Comparison to applet results:

(j) Does the formula for the standard deviation presented in (i) consider the size of the population? Is this consistent with what you observed in (g) and (h)?

(k) When  $n = 20$  does the distribution of sample proportions appear to be well-modeled by a normal distribution? (Check the **Overlay Normal Distribution** box to help you decide.)

In fact, the distribution of sample proportions when taking simple random samples from a population should remind you very much of the distribution of sample proportions when sampling from a process, such as the Reese's Pieces investigation. In fact, the same “Central Limit Theorem” also result holds in this case of sampling from a population, with one additional caveat about the relative sizes of the sample and population, as summarized below:

**Key Result:** The [Central Limit Theorem for a sample proportion](#) states that when drawing random samples from a large but finite population with a large enough sample size, then the *sampling distribution* of the sample proportion  $\hat{p}$  will be well modeled by a normal distribution with mean equal to  $\pi$ , the population proportion of successes, and standard deviation equal to  $SD(\hat{p}) = \sqrt{\pi(1-\pi)/n}$ .

We consider the sample size large enough if  $n\pi \geq 10$  and  $n(1-\pi) \geq 10$ .

We consider the population size ( $N$ ) large if it is more than 20 times the size of the sample.

If the population size is not large, we would use a different standard deviation formula that incorporates a “finite population correction factor.” Probability sampling other than simple random sampling would also require calculating a different standard deviation. You can learn more appropriate techniques in a course on sampling design and methods.

This result means that the normal-based procedures for making inferences about an underlying *process* also apply for making inferences from sample data to a larger *population*, provided that the sample is selected randomly from the (large) population.

**Discussion:** Once again you see the fundamental principle of *sampling variability*. Fortunately, this variability follows a predictable pattern in the long run. You should see that the average of the sample proportions is approximately equal to the population proportion,  $\pi = 0.466$ , as we would expect because the simple random sampling method is unbiased. [The mean of your empirical sampling distribution may not be exactly equal to  $\pi$  as you only took 1000 samples instead of all possible samples of size 5 or 20, but we are no longer consistently overestimating the population parameter as we would with the nonrandom samples.] The second key advantage of randomly sampling is that this predictable pattern will allow us to estimate how far the sample statistic is likely to fall from the population parameter and when we might have a surprising value for the sample statistic. Of course our level of surprise will depend on the sample size because you also saw that the amount of sampling variability in the sample proportion decreases when we increase the sample size. Sample proportions that are based on larger samples will tend to fall even closer to the population proportion  $\pi$  and thus there is less variability among the sample proportions. So first, select *randomly* to avoid bias, and then if we can *increase* the

sample size, this will improve the *precision* of the sample results. Once we know the precision of the sample results in repeated samples, then we can decide whether any one particular sample result can be considered statistically significant or unlikely to happen by chance – the random sampling process – alone. A small p-value does not guarantee that the sample result did not happen just by chance, but it does allow you to measure how unlikely such a result is.

In the next investigation you will consider a method for computing an “exact” p-value, taking the population size into account. However, in many situations we don’t know the size of the population, just that it is large, and instead we will proceed directly to the normal-based method, as long as the technical conditions are met.

### Practice Problem 1.14

(a) Which of the following are advantages of studies with a larger sample size? (*Check all that apply*)

- Better represent the population (reduce sampling bias)
- To more precisely estimate the parameter
- To decrease sampling variability
- To make simulation results more precise
- Other?

(b) In conducting a simulation analysis, why might we take a larger number of samples? (*Check all that apply*)

- Better represent the population (reduce sampling bias)
- To more precisely estimate the parameter
- To decrease sampling variability
- To make simulation results more precise
- Other?

**Investigation 1.15: Freshmen Voting Pattern**

A student project group (Norquest, Bayer, & McConville, 2004) wanted to determine whether there was a preference for the primary presidential candidates (George Bush and John Kerry) among first-year students at their university. They believed that these younger students would be more liberal and more inclined to vote for the democratic candidate (Kerry). They took a sample of 30 students from the 705 first-years at their school by assigning the first-year residence halls a number between 1 and 5 (inclusive) and rolling a die to select a residence hall. Then from the selected residence hall they took every seventh room to be part of the sample. The survey was distributed by going from room to room with the surveys, giving each resident of the room a copy of the survey to fill out, and immediately collecting the surveys after they were completed. If one of the residents was not in the room at the time of the first visit, the group repeatedly returned to the room until the person was contacted. The surveys were anonymous and the group members did not look at the completed surveys until they were randomly scrambled. The survey contained three questions, the first question was whether the student planned to vote for Kerry or Bush, and the other two questions concerned other social issues. For half of the surveys Kerry was listed first and for half of the surveys Bush was listed first.

(a) Identify the observational units and response variable in this study. Is the response variable quantitative or categorical? How you are defining a “success”?

(b) Identify the sample, the population, and the sampling frame used in this study.

sample

population

sampling frame

(c) Explain why this sampling method is *not* a simple random sample but can still be considered a random sample and likely to be representative of the population.

**Definition:** [Non-sampling errors](#) can occur even after we have a randomly selected sample. They are not associated with the sampling process, but rather with sources of bias that can arise after the sample has been selected.

Sources of non-sampling errors in surveys include the word choice in survey questions, dishonest or inaccurate responses by respondents due to sensitive questions, faulty memory, the order in which questions appear, a leading tone, and the appearance of the interviewer.

- (d) Identify some precautions taken by these students to avoid *nonsampling* errors.

In this study the sample results were that 22 of the 30 first-years planned to vote for John Kerry, whereas 8 planned to vote for George Bush.

- (e) Construct a bar graph to display the results for this sample and write a one-sentence summary of what the graph reveals.

- (f) Define the parameter of interest in this study. What symbol would we use to represent this value?

As in Investigation 1.13, we are considering these data to be a random sample from a finite population (not a simple random sample but we will ignore that distinction for now). However, we don't have access to the population. We can still predict how the outcomes from many different random samples will behave by using a probability model. That is, we want to know the random behavior of  $X$ , the number of Kerry voters in the sample.

- (g) Explain why we *cannot* treat  $X$  as a binomial random variable.

When sampling from a finite population, the appropriate probability model is a *hypergeometric distribution* (see the Probability Detour). This gives us

$$P(X = k) = \frac{C(M, k)C(N - M, n - k)}{C(N, n)}$$

where  $N$  is the population size,  $M$  is the number of successes in the population, and  $n$  is the sample size.

- (h) Suppose 352 first-year students at this school (essentially 50% of the population) plan to vote for Kerry. Use the hypergeometric distribution to calculate the probability of observing 22 or more students in a random sample of 30 students favoring Kerry. Would the sample result obtained by these students be surprising?

## Probability Detour – Hypergeometric Random Variables

To be a [hypergeometric random variable](#), a random process must have the following properties:

- Observations are drawn without replacement from a population of  $N$  objects.
- There are two distinct types of objects in the population,  $M$  successes and  $N - M$  failures.

The main distinction between a hypergeometric random variable and a binomial random variable is the trials are no longer independent. The probability of drawing a success for the first object is  $M/N$ . But if we do draw a success, then the probability of success for the second object is  $(M - 1)/(N - 1)$ . If we replaced the item, then we would be back to a binomial process.

So to calculate hypergeometric probabilities, we need to consider another probability rule.

- When outcomes are equally likely, then the probability of an event equals the number of ways for the event to happen divided by the number of possible outcomes.

For example, there are 268 words in the Gettysburg Address, 123 contain the letter  $e$  (about 46%) and 145 do not. Using counting rules, there are  $C(268, 5) = 11,096,761,368$  different samples of 5 words that we could select from the Gettysburg Address. But if we want to select one  $e$ -word and four non- $e$ -words, there are  $C(123, 1) \times C(145, 4) = 2,172,945,060$  such samples. Thus we find  $P(\text{one } e\text{-word})$ ,

$$P(X = 1) = 2,172,945,060 / 11,096,761,368 \approx 0.1958.$$

In general, the probability of obtaining  $k$  successes from a population with  $M$  successes and  $N - M$

failures is  $P(X = k) = \frac{C(M, k)C(N - M, n - k)}{C(N, n)}$  where  $C(a, b) = \binom{a}{b} = \frac{a!}{b!(a - b)!}$

The expected value of the hypergeometric random variable is  $E(X) = (M/N) \times n$  and the standard deviation is  $\sqrt{n \times (M/N) \times (N - M)/N \times (N - n)/(N - 1)}$ . (Compare these to binomial with  $\pi = M/N$ .)

### Binomial Approximation to Hypergeometric Distribution:

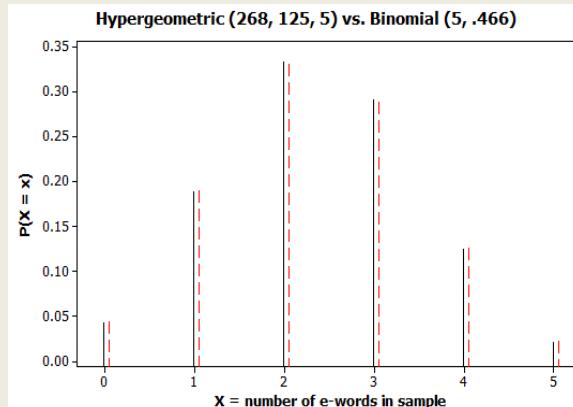
If we instead used the binomial distribution to approximate this probability, we would find

$$P(X = 1) = C(5, 1)(125/268)(143/268)^4 \approx 0.1890.$$

These probability calculations work out to be very similar! This happens when the finite population we are sampling from is large compared to our sample size (e.g.,  $N > 20 \times n$ ).

This means we can continue to use the binomial distribution to determine such probabilities and the expected number of successes  $5(0.466) \approx 2.33$  with  $SD(X) \approx 1.115$  successes.

And when the sample size is large relative to the probability of success, we can approximate the binomial with the normal distribution.



**Technology Detour – Calculating Hypergeometric Probabilities**

In R, the `iscamhyperprob` function takes the following inputs:

- $k$ , the observed value of interest (or the difference in conditional proportions, assumed if value is less than one, including negative)
- $total$ , the total number of observations in the population
- $succ$ , the overall number of successes in the population
- $n$ , the sample size
- $lower.tail$ , a Boolean which is TRUE or FALSE

For example: `iscamhyperprob(k=22, total=705, succ=352, n=30, lower.tail=FALSE)`

(i) The student project group suspected that about 2/3 of the first-year students at this school are planning to vote for Kerry. If 470 first-year students at this school plan to vote for Kerry (2/3 of this population), would the sample result obtained by these students be surprising? Explain.

(j) When the population size is much larger than the sample size, we can use the binomial distribution to approximate the hypergeometric distribution. This approximation is generally considered valid when  $N$ , the population size is 10 to 20 times larger than  $n$ , the sample size. Would you consider this approximation valid for this study? Explain.

**Discussion:** The hypergeometric distribution is able to tell us if we listed all possible samples of 30 students from this population, how many would be at least as extreme as the observed sample. In many sampling situations, our population size is large enough that the binomial distribution gives a very reasonable approximation to the exact p-value. For this reason, the binomial p-value is much more commonly used than the hypergeometric p-value.

### Study Conclusions

Only about 0.68% of random samples from a population where the two candidates were equally preferred would have 22 or more successes (Kerry voters) by random chance alone. This gives us very strong evidence that more than 50% of this population was in favor of Kerry. In fact, we also found that 2/3 is a plausible (believable) value for the population proportion in favor of Kerry. We applied the hypergeometric distribution because we had a small population rather than an infinite process, but the population was large enough that we could have applied the binomial distribution instead.

### Practice Problem 1.15

Reconsider the voter pattern study. Is 60% a plausible value for the percentage of students at this university who were in favor of Kerry? Justify your answer.

### Finite Population Correction Factor

The binomial approximation to the hypergeometric distribution is very convenient for us as typically we don't know the exact size of the population. But when we do have a small population, it's important to realize that the observations are not independent, and the formula we found for the standard deviation of sample proportions  $\sqrt{\pi(1-\pi)/n}$  may not be a very good approximation. Instead, we want to use the

$$\text{standard deviation for a hypergeometric random variable } \text{SD}(\hat{p}) = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Notice that the only difference between these two formulas is the first factor:  $\sqrt{\frac{N-n}{N-1}}$ . This is referred

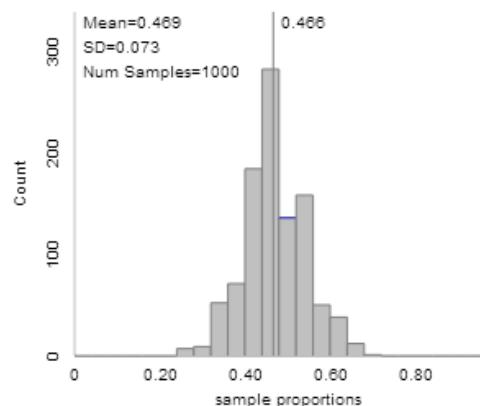
to as a "finite population correction factor." It tells us the factor by which we should *decrease* the measure of sample to sample variability in the statistic. This multiplier does approach one as  $N$  approaches infinity, or in other words, as our population becomes much, much larger than our sample.

For the Gettysburg example with  $n = 40$ , this would violate our "large population" technical condition. So we would calculate

$$\sqrt{\frac{268-40}{268-1}} \sqrt{\frac{0.466(1-0.466)}{40}} = 0.0729 \text{ instead of}$$

$$\sqrt{\frac{0.466(1-0.466)}{40}} = 0.0799.$$

This standard deviation is a more accurate measure of the sample to sample variability when the population size is not very large.



**Investigation 1.16: Teen Hearing Loss**

Shargorodsky, Curhan, Curhan, and Eavey (2010) examined hearing loss data for 1771 participants from the National Health and Nutrition Examination Survey (NHANES, 2005-6), aged 12-19 years. (“NHANES provides nationally representative cross-sectional data on the health status of the civilian, non-institutionalized U.S. population.”) In this sample, 333 were found to have at least some level of hearing loss. News of this study spread quickly, many blaming the prevalence of hearing loss on the higher use of ear buds by teens. At MSNBC.com (8/17/2010), Carla Johnson summarized the study with the headline “1 in 5 U.S. teens has hearing loss, study says.” We will first decide whether we consider this an appropriate headline.

- (a) Define the population and parameter of interest in this study (in symbols and in words).

Population:

Parameter:

- (b) State the null and alternative hypotheses for deciding whether these data provide convincing evidence against this headline (in symbols and in words).

**Discussion:** A consequence of the expansion of the Central Limit Theorem we made in the previous investigation is that we can use all the same normal-based  $z$ -procedures when we have a large enough sample size and have taken a random sample from a large population.

- (c) Under the null hypothesis, is the normal model appropriate for the distribution of sample proportions based on this sample size and on this population size? Justify your answers numerically.

- (d) Use technology to carry out a  $z$ -test of the hypotheses you specified in (b). Report the test statistic and p-value, and confirm how you would calculate the test statistic by hand.

(e) Does this p-value provide convincing evidence ( $\alpha = 0.05$ ) against the headline? Justify your answer (in context) and explain the reasoning behind your evidence or lack thereof.

(f) Calculate and interpret a 95% confidence interval. Is this interval consistent with your test decision?

(g) Do you feel comfortable with generalizing the findings from your test and confidence interval to the population of all American teens in 2005-06? Explain.

### Study Conclusions

We have no reason to doubt that the NHANES sample will be representative of the larger population of U.S. teens. Although we are not told how they selected the sample, they claim it was nationally representative and presumably involved a probability sampling method. The population size is quite large compared to the sample size of nearly 1,800 teens, so that condition for using the normal model for the sampling distribution of the sample proportions is met. The sample size is also very large relative to the hypothesized population proportion of 0.20 ( $0.20 \times 1771 = 354.2$  and  $0.80 \times 171 = 1416.8$  both easily exceed 10). Therefore calculations from the binomial probability model and the normal probability model will be quite similar. Under the null hypothesis that 20% of U.S. teens have some hearing loss, the sampling distribution of the sample proportion will be approximately normal, with a mean of 0.20 and a standard deviation of 0.0095. Therefore our sample proportion,  $\hat{p} = 333/1771 = 0.188$  lies 1.26 standard deviations below the hypothesized population proportion, yielding a two-sided p-value of 0.2068 (TBI applet). This p-value is not small and does not provide convincing evidence against the claim that  $\pi = 0.20$ . We would get sample proportions as or more extreme as 0.188 (meaning as far from 0.20) in about 21% of random samples from a population where 20% of teens have some hearing loss. Therefore the headline seems appropriate. The confidence interval (TBI applet) tells us that we are 95% confident that between 17.0% and 20.6% of all U.S. teens have some form of hearing loss. (So keep in mind that 0.20 is just one plausible value.) It is difficult to evaluate these values without knowing whether this percentage is increasing over time which we will investigate later.

## Summary of One Proportion z-Procedures for Proportion of Large Population

Let  $\pi$  the (unknown) population proportion of successes and given that you have a representative sample from the population of interest.

### Test of $H_0: \pi = \pi_0$

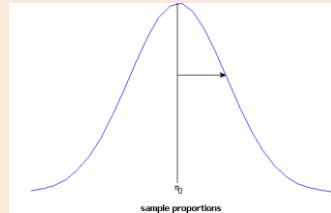
*Test statistic:*  $z_0 = (\hat{p} - \pi_0) / \sqrt{\pi_0(1 - \pi_0)/n}$

*p-value:*

If  $H_a: \pi > \pi_0$ , the p-value is  $P(Z \geq z_0)$ .

If  $H_a: \pi < \pi_0$ , the p-value is  $P(Z \leq z_0)$ .

If  $H_a: \pi \neq \pi_0$ , the p-value is  $2P(Z \geq |z_0|)$



### C% confidence interval for $\pi$ :

- *Wald Interval:*  $\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})/n}$  where  $-z^*$  is the  $100 \times (1 - C)/2$ th percentile of the standard normal distribution.

*Technical conditions:*  $n \hat{p} \geq 10$  and  $n(1 - \hat{p}) \geq 10$  (which means there are at least 10 successes and 10 failures in the sample) and population size  $\geq 20n$ .

- *Adjusted Wald Interval:*  $\tilde{p} \pm z^* \sqrt{\tilde{p}(1 - \tilde{p})/\tilde{n}}$  where  $\tilde{p} = (X + 0.5z^{*2})/(n + z^{*2})$  and  $\tilde{n} = n + z^{*2}$  (For 95% confidence, add two successes and two failures.)

*Technical conditions:* population size  $\geq 20n$

### In Theory-Based Inference applet: One Proportion

- Specify the sample size and either the sample count or the sample proportion
  - Remember to adjust these values if performing the Adjusted Wald interval
- Check the Test of Significance box and/or the Confidence Interval box
  - Specify the value of  $\pi_0$ , use the < button to change the direction of the alternative
  - Specify the confidence level

### In R: iscamonepropztest(observed, n, hypothesized, alternative, conf.level)

- Specify the number of successes or  $\hat{p}$ ,  $n$ ,  $\pi_0$ , alternative, and confidence level (in that order)
  - Remember to adjust inputs if performing the Adjusted Wald interval
- For the alternative, choose “two.sided” or “less” or “greater” (with the quotes)
- If no alternative is specified, be sure to label the confidence level (conf.level)

**Investigation 1.17: Cat Households**

A national survey of over 47,000 households in 2007 found that 32.4% of American households own a pet cat or vice versa ([2007 U.S. Pet Ownership & Demographics Sourcebook](#)).

- (a) Identify the observational units and variable of interest.

Observational units:

Variable:

- (b) Is 32.4% a parameter or a statistic? Indicate the symbol used to represent it.

- (c) Conduct a test of whether the sample data provide evidence that the population proportion who own a pet cat differs from one-third. State the hypotheses, and report the test statistic and p-value. State your test decision at the  $\alpha = 0.01$  level, and summarize your conclusion in the context of this study.

- (d) Explain why the p-value turns out to be so very small, when the sample proportion of households with a cat seems to be quite close to one-third.

- (e) Produce a 99% confidence interval (CI) for the population proportion who own a pet cat. Interpret this interval.

- (f) Is the confidence interval consistent with the test decision? Explain.

- (g) Do the sample data provide *very* strong evidence that the population proportion who own a pet cat is not one-third? Explain whether the p-value or the CI helps you to decide.

- (h) Do the sample data provide strong evidence that the population proportion who own a pet cat is *very* different from one-third? Explain whether the p-value or the CI helps you to decide.

**Discussion:** Keep in mind the difference between *statistical significance* and *practical significance*. With large sample sizes, sample proportions will vary little from sample to sample, and so even small differences (that may seem minor to most of us) will be statistically significant. Saying that a sample result is unlikely to happen by chance (and therefore is statistically significant) is not the same as saying the result is important or even noteworthy (practically significant), depending on the context involved.

### Investigation 1.18: Female Senators

Suppose that an alien lands on Earth, notices that there are two different sexes of the human species, and sets out to estimate the proportion of humans who are female. Fortunately, the alien had a good statistics course on its home planet, so it knows to take a sample of human beings and produce a confidence interval. Suppose that the alien happened upon the members of the 2015 U.S. Senate as its sample of human beings, so it finds 20 women and 80 men in its sample.

(a) Use this sample information to form a 95% confidence interval for the actual proportion of all humans who are female.

(b) Is this confidence interval a reasonable estimate of the actual proportion of all humans who are female?

(c) Explain why the confidence interval procedure fails to produce an accurate estimate of the population parameter in this situation.

(d) It clearly does not make sense to use the confidence interval in (a) to estimate the proportion of women on Earth or even the U.S., but does the interval make sense for estimating the proportion of women in the 2015 U.S. Senate? Explain your answer.

**Discussion:**

- First, statistical tests and confidence intervals do not compensate for the problems of a biased sampling procedure. If the sample is collected from the population in a biased manner, the ensuing confidence interval will be a biased, and potentially misleading, estimate of the population parameter of interest.

- A second important point to remember is that confidence intervals and significance tests use *sample statistics* to estimate *population or process parameters*. When the data at hand constitute the entire population of interest, then constructing a confidence interval from these data is meaningless. In this case, you know precisely that the proportion of women in the population of the 2015 U.S. Senators is 0.20 (exactly! no margin-of-error!), so it is senseless to construct a confidence interval from these data.

**Example 1.1: Predicting Elections from Faces?**

*Try these questions yourself before you use the solutions following to check your answers.*

Do voters make judgments about a political candidate based on his/her facial appearance? Can you correctly predict the outcome of an election, more often than not, simply by choosing the candidate whose face is judged to be more competent-looking? Researchers investigated this question in a study published in *Science* (Todorov, Mandisodka, Goren, and Hall, 2005). Participants were shown pictures of two candidates and asked who has the more competent looking face. Researchers then predicted the winner to be the candidate whose face was judged to look more competent by most of the participants. For the 32 U.S. Senate races in 2004, this method predicted the winner correctly in 23 of them.

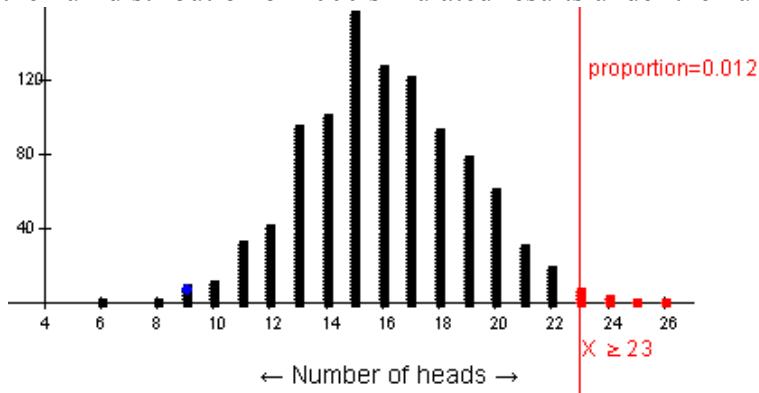
- (a) In what proportion of the races did the “competent face” method predict the winner correctly?
- (b) Describe (in words) the null model to be investigated with this study.
- (c) Describe how you could (in principle) use a coin to produce a simulation analysis of whether these data provide strong evidence that the “competent face” method would correctly predict the election winner more than half the time. Include enough detail that someone else could implement the full analysis and draw a reasonable conclusion.
- (d) Use the One Proportion Inference applet to conduct a simulation (using 1000 samples) addressing the question of whether the researchers’ results provide strong evidence in support of the researchers’ conjecture that the “competent face” method would correctly predict the election winner more than half the time. Be sure to report (and indicate on the dotplot), the p-value from the simulation.
- (e) Write a paragraph, as if to the researchers, describing what your simulation analysis reveals about whether the data provide strong evidence in support of their conjecture.

- (f) These researchers also predicted the outcomes of 279 races for the U.S. House of Representatives in 2004. The “competent face” method correctly predicted the winner in 189 of those races. Do you predict this p-value will be larger or smaller than in (d)? Explain your reasoning.

## Analysis

- (a) This method predicted the winner correctly in  $23/32 \approx 0.719$  of the races.  
 (b) The null model is that the “competent face” method is not useful for predicting election winners, so it would work in only 50% of all races in the long run.  
 (c) You would flip a fair coin 32 times, once for each of the election races. Keep track of the number of heads, representing races for which the method correctly predicted the winner. Repeat simulating 32 flips a large number of times. Determine the proportion of samples in which 23 or more coin tosses resulted in heads. If this proportion is very small, conclude that the competent face method really does predict the winner correctly more than half the time.

- (d) Here is a graph of the null distribution of 1000 simulated results under the null model:



Based on these 1000 samples, the approximate p-value is  $12/1000 = 0.012$ . (Your approximate p-value will probably be different but close to this. The exact Binomial p-value equals 0.01.)

- (e) This analysis reveals that it would be quite surprising to correctly predict 23 or more of 32 election races if the “competent face” method were no better than flipping a coin. More specifically, such an extreme result would only occur about 1% of the time by random chance. So, the data provide fairly strong evidence in support of the researchers’ conjecture that the “competent face” method works more than half the time.

- (f) The observed proportion of successes is now  $189/279 \approx 0.677$ . This is not quite as extreme (far from 0.5) as the first study, but the sample size is much larger. For this reason, you probably expect the p-value to be even smaller. (Indeed, the p-value now dips below 0.001.)

**Example 1.2: Cola Discrimination**

*Try these questions yourself before you use the solutions following to check your answers.*

A teacher doubted whether his students could distinguish between the tastes of different brands of cola, so he presented each of his 21 students with three cups. Two cups contained one brand of cola, and the third cup contained a different brand. Which cup contained which brand was randomly determined for each student. Each student was asked to identify which cup contained the cola that was different from the other two. It turned out that 12 of the students successfully identified the “odd” cola.

- (a) Does this result provide strong evidence that these students do better than guessing in discriminating among the colas? Address this question with an appropriate test of significance, including a statement of the hypotheses and a p-value calculation or estimation. Be sure to clarify which procedure you used to determine the p-value and why. Summarize your conclusion, and explain the reasoning process by which it follows.
- (b) Calculate a 95% confidence interval based on these sample data. Clearly define the parameter that this interval estimates, and interpret the interval.
- (c) Would this teacher be convinced that his students do better than guessing if he uses the 0.05 significance level?

(d) Describe what Type I and Type II errors mean in this situation.

(e) Suppose the students want to redo the study with 100 students. Would a normal approximation be valid in this case? Justify your answer.

(f) Using the 0.05 significance level and a sample size of 100 students, calculate the power of the test when the underlying probability equals 0.5, and interpret your result.

[Hint: Remember our two-step process: 1) Find the rejection region under the null hypothesis, now using the normal distribution; 2) Determine the probability of being in the rejection region under an alternative value of the parameter.]

(g) If the underlying probability of identifying the odd soda equals  $2/3$ , will the power calculated in (f) be larger or smaller in this case? Explain without performing any calculations.

(h) If the teacher uses a 0.01 significance level (with an alternative probability of 0.50), will the power calculated in (f) be larger or smaller in this case? Explain without performing any calculations.

## Analysis

(a) We can define  $\pi$  to be the probability that these students correctly identify the odd soda. (In other words, if this group of students were to repeat this process under identical conditions indefinitely,  $\pi$  represents the long-term fraction that they would identify correctly.) The null hypothesis asserts that the students are just guessing, which means that their success probability is one-third ( $H_0: \pi = 1/3$ ). The alternative hypothesis is that students do better than guessing, which means that their success probability is greater than one-third ( $H_a: \pi > 1/3$ ).

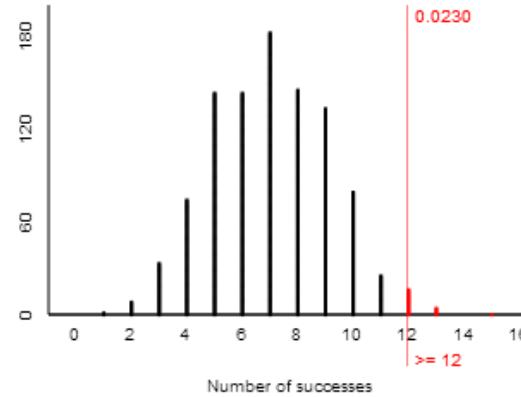
Under the null hypothesis that the students are just guessing among the three cups,  $X$  has a binomial distribution with parameters  $n = 21$  and  $\pi = 1/3$ . The normal distribution would probably not be valid here as we do not satisfy the conditions:  $n \times \pi = 21(1/3) = 7 < 10$ . We could simulate observations from this Binomial process using the One Proportion Inference applet, and see how often we observe 12 or more correct identifications just by chance:

Probability of success ( $\pi$ ):	<input type="text" value="0.3333"/>
Sample size ( $n$ ):	<input type="text" value="21"/>
Number of samples:	<input type="text" value="1000"/>
As extreme as	<input type="button" value="≥"/> <input type="text" value="12"/> <input type="button" value="Count"/>

Proportion of samples:  
 $23 / 1000 = 0.0230$

Exact Binomial

$P(X \geq 12) = 0.0212$

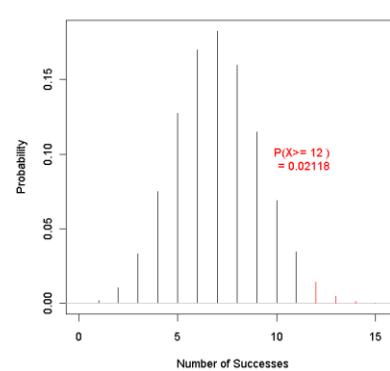


Or we could calculate the exact Binomial probability, either with this applet or with

```
> iscabinomprob(k=12, n=21, prob=.3333,
lower.tail=FALSE)
Probability 12 and above = 0.02117713
```

or using `iscabinomtest`:

```
> iscabinomtest(observed=12, n=21,
hyp=.3333, alternative="greater")
```



Exact Binomial Test

```
Data: observed successes = 12, sample size = 21, sample proportion = 0.5714
Null hypothesis      : pi = 0.3333
Alternative hypothesis: pi > 0.3333
p-value: 0.021177
```

Notice the normal approximation does not provide a great estimate of this p-value

```
> iscamonepropztest(observed=12, n=21, hyp=.3333,
alternative="greater")
```

One Proportion z test

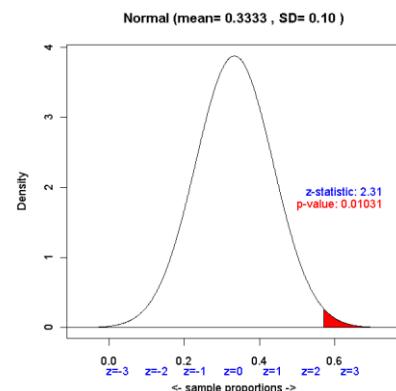
```
Data: observed successes = 12, sample size
= 21, sample proportion = 0.5714
```

Null hypothesis :  $\pi = 0.3333$

Alternative hypothesis:  $\pi > 0.3333$

z-statistic: 2.31

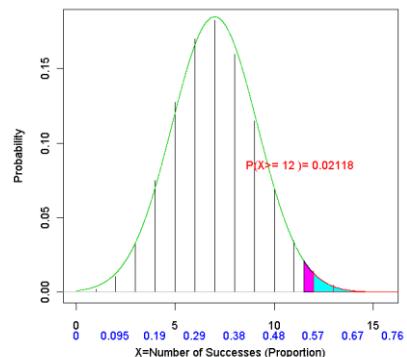
p-value: 0.01031



But the normal approximation does improve with the continuity correction

```
> iscambinomnorm(k=12, n=21, prob=.3333,
direction="above")
binomial: 0.02118
normal approx: 0.01031
normal approx with continuity: 0.01860
```

Binomial ( $n = 21$ ,  $\pi = 0.3333$ ), Normal(mean = 0.3333, SD = 0.10)



This p-value reveals that if the students were just guessing, there's only about a 2% probability of “by chance” getting 12 or more correct identifications among 21 trials. In other words, if we repeated this study over and over, **and if** students were just guessing each time, then a result this favorable would occur in only about 2% of the studies. Because this probability is quite small, we have fairly strong evidence that these students’ process in fact is better than guessing in discriminating among the colas (i.e., that  $\pi > 1/3$ ). Because the sodas were randomly placed in the cups and (presumably) the teacher kept other variables (e.g., temperature, age) constant, this study attempted to isolate the taste and appearance of the sodas as the sole reasons for their selection.

(b) If we calculate the default binomial interval we find (e.g., in R):

```
> iscambinomtest(12, 21, conf.level=95)
```

95 % Confidence interval for pi: ( 0.34021 , 0.7818 )

So we are 95% confident that the underlying probability of a correct identification for these students is between 0.340 and 0.782.

(c) Yes, the p-value is less than 0.05, so with that significance level, the teacher would be convinced that his students do better than guessing in discriminating among the colas. Notice also that 1/3 is not captured in the 95% confidence interval.

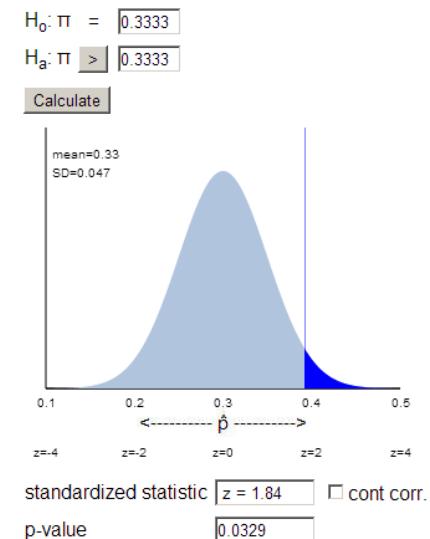
(d) A Type I error would mean that these students are actually just guessing, but we erroneously conclude that they do better than guessing. A Type II error would mean that the students are not just guessing, but we do not consider the evidence strong enough to conclude that.

(e) Now the normal approximation should be valid. If we assume  $\pi = 1/3$ , then  $n\pi = 1000(1/3) \approx 33.3 > 10$  and  $n(1 - \pi) = 100(2/3) \approx 66.7 > 10$ . (If  $\pi$  is larger than  $1/3$  as we suggested above, this approximate is even more valid.)

(f) If we want to test  $H_0: \pi = 1/3$  vs.  $H_a: \pi > 1/3$  for a sample size of  $n = 100$ , then the rejection region will be the values of  $\hat{p}$  so that

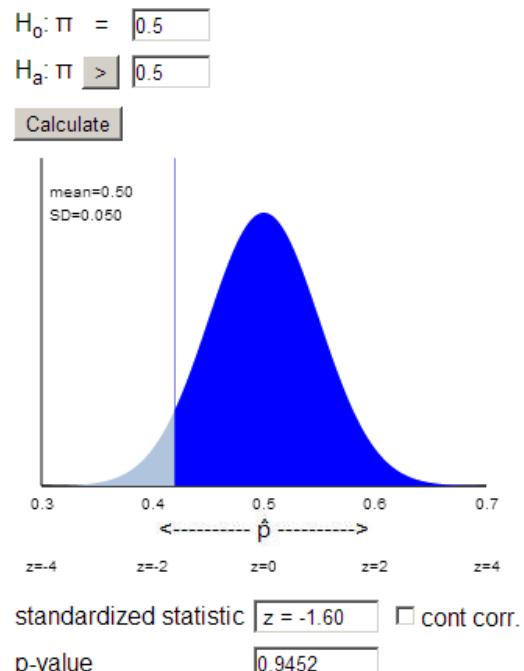
$P(\hat{p} > \text{cut-off}) \leq 0.05$  when  $\pi = 1/3$ . Because we are comfortable using the normal approximation with this sample size, we can use the Theory-Based Inference applet to find this cut-off (enter the hypothesized value, the direction of the alternative, the sample size, then change the count input until the p-value dips below 0.05).

This says a  $\hat{p}$  of 0.420 or larger would lead us to reject this null hypothesis that  $\pi = 0.5$  at the 5% level of significance.

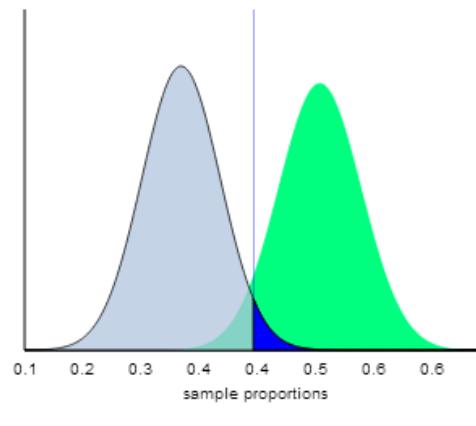
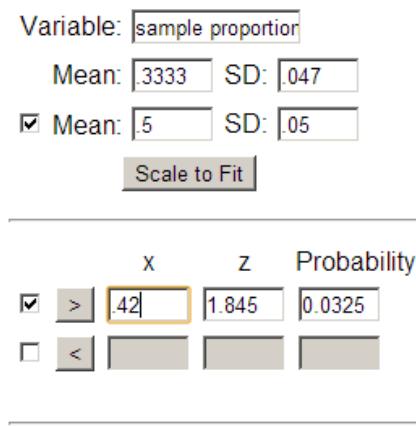


So now we need to find the probability  $P(\hat{p} \geq 0.420)$  when the underlying process probability is actually 0.5. When  $\pi = 0.5$ , we have an approximately normal distribution with mean equal to 0.5 and standard deviation equal to  $\sqrt{0.5(0.5)/100} = 0.05$ . Using the TBI applet, We can enter 0.5 as the hypothesized probability and 0.420 as the observed sample proportion with a  $>$  alternative. This gives us an appropriate power of 0.9452.

Thus, the probability of rejecting  $H_0: \pi = 1/3$  when  $n = 100$  and  $\pi$  is actually 0.5 equals 0.9452. Therefore, in about 95% of samples from a process with  $\pi = 0.5$  we will correctly reject that  $\pi = 1/3$ .

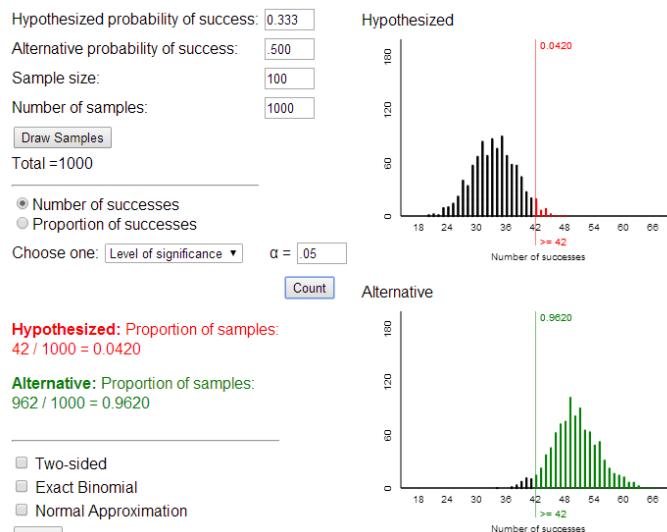


Showing these curves on the same graph:



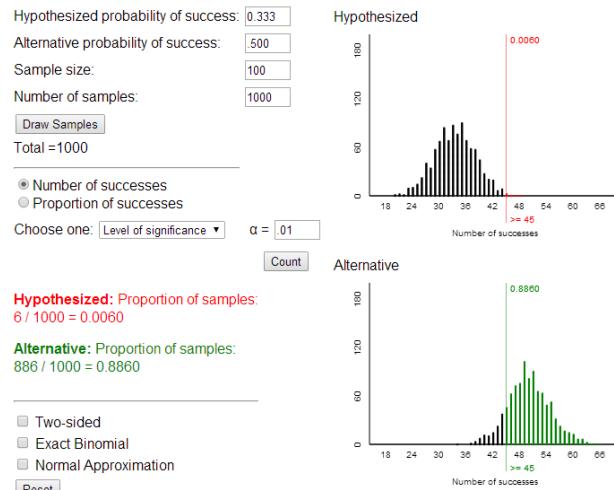
Power is the proportion of the green area that lies to the right of the blue line.

Confirming our results with the Power Simulator applet:



(g) If the underlying process probability is even larger at  $\pi = 2/3$ , this will shift the alternative distribution to the right and even more of the distribution will lie in the rejection region. Our power will increase as it will be even easier for us to obtain a sample that will convince us that the hypothesized value of 1/3 is incorrect.

(h) If the level of significance is lowered from 0.05 to 0.01, this will move the line to the right. We will need a larger sample proportion to convince us and this will slightly reduce our power (less than 0.90).



**Example 1.3: Seat Belt Usage**

Try these questions yourself before you use the solutions following to check your answers.

Every year since 1983 the Harris Poll has measured and reported the key lifestyle characteristics and behaviors which are known to have a major impact on health, disease, injury and life expectancy. One such survey was conducted by telephone with a nationwide sample of 1,005 adults between March 11 and 16, 1998. In this survey, 77% of adults claimed to always wear seat belts when they are in the front seat of a car.

- (a) Identify the observational units and population of interest.
- (b) Define the relevant population parameter in words.
- (c) Use this sample data to construct a 95% confidence interval for the population parameter *using two different methods*. Compare the results of your methods.
- (d) Explain what is meant by the phrase “95% confidence” for these procedures.
- (e) Suggest a reason we might be skeptical that this confidence interval captures the proportion of adults who use a seat belt while in the front seat of a car.
- (f) Suggest some ways to guard against non-sampling errors in this study. That is, how might you collect the data to get a more accurate estimate of the proportion of drivers that actually use a seat belt.

## Analysis

- (a) The observational units are U.S. adults and the population consists of all U.S. adults (in 1998).
- (b) The parameter of interest is the proportion of all U.S. adults that (claim to) always wear a seat belt when they are in the front seat of a car.
- (c) We can use several methods because the population (all adults) is much larger than  $20 \times 1005 = 20,100$ .

**Binomial CI** in R with `iscbinomtest`  
 $(X = 0.77 \times 1005 = 774)$

**Wald interval** (TBI applet)  
Valid because  $774$  and  $231 > 10$ .

**Plus Four** (TBI applet with  $776$  and  $1009$ )

**Score** in R with `prop.test`

95 percent confidence interval:  
 $0.7428691 \text{ } 0.7958367$

confidence level  %  
 $(0.7441, 0.7961)$

confidence level  %  
 $(0.7431, 0.7951)$

95 percent confidence interval:  
 $0.7426225 \text{ } 0.7955795$

The intervals are very similar to each other as we would expect with the large sample size, though the binomial interval is a bit wider than the others.

These intervals tell us that we are 95% confident that between 74% and 80% of drivers (claim to) always wear their seat belt. We are willing to generalize this result to all adult Americans because Harris conducted a nationwide poll, presumably using some random selecting techniques.

- (d) If Harris Poll was to repeatedly take random samples of 1,005 adults and calculate a confidence interval for each one, then over thousands and thousands of such samples, we'd expect roughly 95% of the intervals to capture the actual value of the population proportion that (claim to) wear a seat belt.
- (e) This is a bit of a loaded question, with a clear expectation for appropriate behavior. People may overestimate their usage, either to not admit to breaking the law or due to faulty memory. We should be somewhat skeptical of such self-reported responses.
- (f) The researchers could observe drivers rather than only asking drivers. In fact, a 2001 study of 612 drivers coming to a convenience store in El Paso, Texas found 61.5% of the drivers actually wearing a seat belt. But when asked whether they always wear their seat belt, 75% of these same drivers said yes. Researchers also worry about the tone used in asking questions and making sure every interviewer uses exactly the same language and demeanor when conducting a phone poll. In surveys with many questions, researchers will also rotate the order of questions to different respondents. Respondents can also be asked about related behaviors or past efforts (e.g., "did you vote last year" instead of "do you plan to vote").

## CHAPTER 1 SUMMARY

In this chapter, you have focused on making inferences based on a representative sample from a random process that can be repeated indefinitely under identical conditions or from a finite population, for a single binary variable. You have learned how to model the chance variation in the binary outcomes that arise from such a sampling process.

- You utilized *simulation* as a powerful tool for assessing the behavior of the outcomes and estimating the likelihood of different results for the sample proportion. In particular, you saw you could estimate the *p-value* of a test to measure how unlikely we are to get a sample proportion at least as extreme as what was observed under certain conjectures (the null hypothesis) about the process or population from which the sample was drawn.
- Then you used the *binomial distribution* to calculate one-sided and two-sided p-values exactly.
- As a third alternative for estimating p-values and confidence intervals, you considered the normal approximation to the binomial distribution (the *Central Limit Theorem* for a sample proportion). In this case, we also found *z-score* values (*test statistics*). These are informative in accompanying p-value calculations to provide another assessment of how unusual an observation is. We often flag an observation as surprising or unusual if the  $|z\text{-score}|$  value exceeds 2.

In each case, when the p-value is small, we have evidence that the observed result did not happen solely due to the random chance inherent in the sampling process (a “statistically significant result”). The smaller the p-value, the stronger the evidence against the null hypothesis.

Still, we must keep in mind that we are merely measuring the strength of evidence – we may be making a mistake whenever we decide whether or not to reject a null hypothesis. If we decide the observed result did not happen by chance and so reject the null hypothesis, there is still a small probability that the null hypothesis is true and that the observed result did occur by chance (the probability of committing a *Type I error*). If we decide the result did happen by chance, there is still a probability that something other than random chance was involved (a *Type II error*). It is important to consider the probabilities of these errors when completing your assessment of a study. In particular, the sample size of the study can influence the probability of a Type II error and the related idea of *power*, which is the probability of correctly rejecting a null hypothesis when it is false.

You began your study of *confidence intervals* as specifying an interval of plausible values for the process probability based on what you observed in the sample. These were the hypothesized parameter values that generated two-sided p-values above the level of significance  $\alpha$ . In other words, they were the parameter values for which your sample result would not be surprising. When the sample size is large (large enough for the normal approximation to be valid), we saw that these confidence intervals have a very convenient form:  $\text{statistic} \pm \text{margin-of-error}$  where  $\text{margin-of-error} = \text{critical value} \times \text{standard error of statistic}$ . The critical value is the number of standard errors you want to use corresponding to a specified confidence level. Keep in mind that the *level of confidence* provides a measure of how reliable the procedure will be in the long-run (which can vary by procedure and sample conditions).

Finally, you saw that this reasoning process holds equally well when the sampling is from a finite population, where the *randomness* in our model comes from the selection of the observational units, not in the observational units’ individual outcomes. Here we took a bit more care to convince ourselves that the sample will be representative of the larger population. This is done by using random mechanisms to

select the sample. These random mechanisms (e.g., simple random sampling) lead to samples that will generally have characteristics mirroring those of the larger population. Although random sampling prevents systematic sampling errors, you still need to worry about non-sampling errors (e.g., wording of a question). Also, there will still be random sampling variability – the characteristics of the sample will vary from sample to sample. Technically we should use the hypergeometric distribution to model the behavior of the statistic. But we saw that if the population is large compared to the size of the sample (e.g., more than 20 times larger), then we can use the same binomial distribution and if the sample size is also large we can use normal-based methods to determine p-values and confidence intervals (as well as power and sample size calculations). The interpretation of the p-value is essentially the same but now applies to the randomness inherent in the sampling process. Also keep in mind that the confidence interval aims to capture the proportion of the population having the characteristic of interest. (Note, when the population is large, these are actually equivalent because the probability of any one randomly selected observational unit being a success will equal the population proportion of successes and we are approximating this as constant for every member of the sample.)

## SUMMARY OF WHAT YOU HAVE LEARNED IN THIS CHAPTER

- The reasoning process of statistical inference
- The terms *parameter* to describe a numerical characteristic of a population or process and *statistic* to describe a numerical characteristic of a sample
- The symbol  $\pi$  to represent the probability of success for a process or the population proportion and  $\hat{p}$  to represent a sample proportion of successes
- The fundamental notion of sampling variability and how to simulate empirical sampling (null) distributions “by hand” (e.g., using cards) and using technology (e.g., with an applet)
- How to estimate and interpret a p-value
- How to use technology to calculate binomial probabilities, as well as exact p-values and confidence intervals (see Example 1.1)
- When and how to apply the Central Limit Theorem for a sample proportion to approximate the binomial distribution (for the values of  $n$  and  $\pi$ ) with a normal distribution (and how to determine the mean and standard deviation of this distribution)
- How to use the Theory-Based Inference applet to estimate p-values and confidence intervals using the normal approximation to the binomial distribution
- The formal structure of a test of significance about a process/population parameter (define the parameter, state null and alternative hypotheses, determine which probability model to use, calculate the p-value as specified by the alternative hypothesis under the assumption that the null hypothesis is true, decide to reject or fail to reject the null hypothesis for the stated level of significance, and state the conclusion in context)
- How to calculate and interpret  $z$ -scores
- What factors affect the size of the p-value
- Type I and Type II errors, Power: what they mean, how their probabilities are determined, and how they are affected by sample size and each other
  - Either through Binomial or Normal calculations (see Example 1.2)
- The idea of a confidence interval as the set of plausible values of the parameter that could have reasonably led to the sample result that was observed
- How to interpret the “confidence level” of an interval procedure
- What factors affect the width, midpoint, and coverage rate of a confidence interval procedure

- The logic and trade-offs behind different confidence interval procedures
- The distinction between statistical and practical significance and how we assess each
- Biased sampling methods systematically over-estimate or under-estimate the parameter; the sampling distribution of a statistic from an unbiased sampling method will center at the value of the parameter of interest
- Random sampling eliminates sampling bias and allows us to use results from our sample to represent the population
- Ways to try to avoid non-sampling errors in a sample survey (see Investigation 1.13 and Example 1.3)

## TECHNOLOGY SUMMARY

- You used applets to explore sampling distributions.
  - The “One Proportion Inference” applet allowed you to explore properties of the binomial distribution. This applet provides both empirical and exact binomial probability calculations. This is similar to what you can do with the “Reese’s Pieces” applet.
  - The “Simulating Power” applet allowed you to compare the distribution under the null hypothesis to the distribution under an alternative hypothesis and consider the probabilities of Type I and Type II errors and how they are controlled.
  - The “Simulating Confidence Intervals” applet allowed you to compare the coverage rates of the Wald, Plus Four/Adjusted Wald, and Score intervals
- You used applets to perform normal probability calculations
  - The “Normal Probability Calculator” applet allowed you to sketch, label, and find areas under a normal curve
  - The “Theory-Based Inference” applet allowed you to carry out a test of significance for a process probability and to find a two-sided confidence interval using the normal approximation to the binomial.
- In R, you learned how to
  - Construct a bar graph using data in a data file
  - Calculate probabilities and quantiles from a binomial distribution (`iscambinomprob`)
  - Calculate exact binomial p-values and confidence intervals using `iscambinomtest` (with summarized data)
  - Calculate the power of a binomial test (`iscambinompower`)
  - Calculate probabilities from a normal distribution (`iscamnormprob`)
  - Calculate quantiles from a normal distribution (`iscaminvnorm`)
  - Calculate one-sample z-test p-values and confidence intervals using `iscamonepropztest`
  - Explore the normal approximation to the binomial and continuity correction (`iscambinomnorm`)

## Choice of Procedures for Analyzing One Proportion

<b>Study design</b>	One binary variable (success/failure), constant probability of success, independence	
<b>Null Hypothesis</b>	$H_0: \pi = \pi_0$	
<b>Parameter</b>	$\pi$ = probability of success	$\pi$ = population proportion
<b>Simulation</b>	Random sample from binomial process	Random sample from a finite population
<b>Exact p-value</b>	Binomial distribution ( $n, \pi_0$ )	Hypergeometric ( $N, M, n$ )
<b>R commands</b>	<code>iscamonepropztest</code> <ul style="list-style-type: none"> <li>• <i>Observed</i> (either the number of successes or sample proportion)</li> <li>• <math>n</math> (sample size)</li> <li>• <i>hypothesized probability</i> (<math>\pi_0</math>)</li> <li>• <i>alternative</i> ("less", "greater", or "two.sided")</li> <li>• <i>Optional: conf.level(s)</i></li> </ul>	<code>iscamhyperprob</code> <ul style="list-style-type: none"> <li>• <math>k</math> (observed number of successes)</li> <li>• <i>total</i> (population size)</li> <li>• <i>succ</i> (hypothesized number of successes in population)</li> <li>• <math>n</math> = sample size</li> <li>• <i>lower.tail</i> (TRUE or FALSE)</li> </ul>
<b>Minitab</b>	Stat > Basic Statistics > 1 Proportion	Graph > Probability Distribution Plot
<b>Can use <math>z</math> procedures if</b>	At least 10 successes and 10 failures in each group	Population size $\geq 20n$ ; At least 10 successes and 10 failures in each group
<b>Test statistic</b>	$z_0 = (\hat{p} - \pi_0) / \sqrt{\pi_0(1 - \pi_0) / n}$	
<b>Confidence interval</b>	<ul style="list-style-type: none"> <li>• Wald: <math>\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}</math></li> <li>• Adjusted Wald: <math>\tilde{p} \pm z^* \sqrt{\tilde{p}(1 - \tilde{p}) / \tilde{n}}</math> where <math>\tilde{p} = (X + 0.5z^{*2}) / (n + z^{*2})</math> and <math>\tilde{n} = n + z^{*2}</math> (For 95% confidence, add two successes and two failures.)</li> </ul>	
<b>R Commands</b>	<code>iscamonepropztest</code> <ul style="list-style-type: none"> <li>• <i>Observed</i> (either the number of successes or sample proportion), <math>n</math> (sample size)</li> <li>• <i>hypothesized probability</i> (<math>\pi_0</math>)</li> <li>• <i>alternative</i> ("less", "greater", or "two.sided")</li> <li>• <i>Optional: conf.level(s)</i></li> </ul>	
<b>Minitab</b>	Stat > Basic Statistics > 1 Proportion Use "normal approximation" under the Options button.	

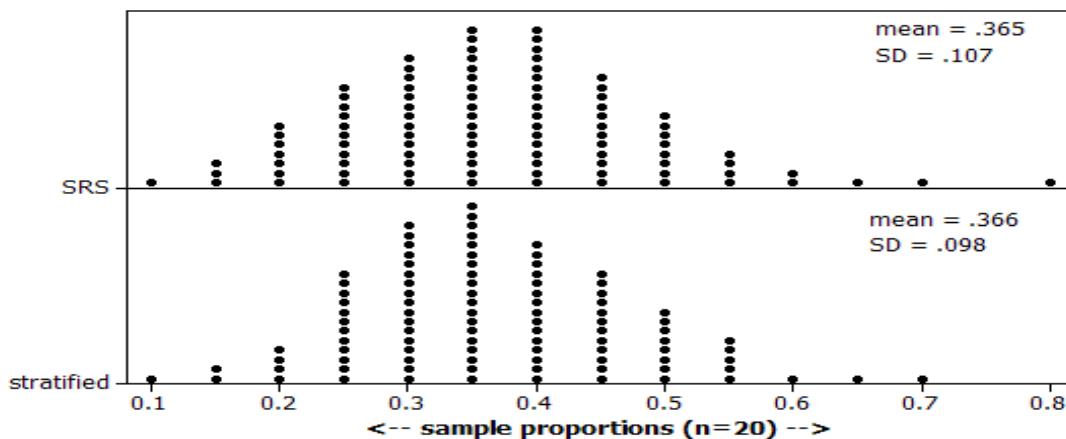
**Quick Reference to ISCAM R Workspace Functions in Chapter 1 (S. Lane-Getaz)**

Probability (or z-score) Desired	Function Name(options)*
<b>Binomial Probability</b>	
Right tail probability of Binomial random variable X, $P(X \geq k   \pi)$	iscabinomprob(k, n, pi, FALSE)
Left tail probability of Binomial random variable X, $P(X \leq k   \pi)$	iscabinomprob(k, n, pi, TRUE)
<b>Exact Binomial Test and Confidence Interval</b>	
Right/Left tail probability of Binomial random variable X, $P(X \geq k   \pi)$	iscabinomtest(k, n, pi, "greater") iscabinomtest(k, n, pi, "less")
Two-tailed probability in Binomial distribution, $P(X \leq k \text{ or } X \geq k_2   \pi)$ w/ method of small P-values	iscabinomtest(k, n, pi, "two.sided")
95% (two-sided) Binomial confidence interval	iscabinomtest(k, n, conf.level = c(90, 95))
<b>Normal Probability</b>	
Right/Left tail probability of Normal random variable X, $P(X \geq x   \mu, \sigma)$	iscannormprob(x, mu, sigma, "above", "x-axis-label") iscannormprob(x, mu, sigma, "below", "x-axis-label")
Two-tailed probabilities in Normal distribution, $P(X \leq x_1 \text{ or } X \geq x_2   \mu, \sigma)$	iscannormprob(x1, mu, sigma, "outside", "x-axis-label", x2)
Probability between two values in Normal distribution, $P(x_1 \leq X \leq x_2   \mu, \sigma)$	iscannormprob(x1, mu, sigma, "between", "x-axis-label", x2)
<b>Inverse Normal Probability</b>	
z-score value with stated probability above/ below	iscaminvnorm(probability, direction = "above") iscaminvnorm(probability, direction = "below")
z-score value with stated probability outside	iscaminvnorm(probability, direction = "outside")
z-score value with stated probability between	iscaminvnorm(probability, direction = "between")
<b>Normal (z-test) Approximation for One Proportion and Confidence Interval (CI)</b>	
Right/Left tail probability of a proportion using Normal z-test, $P(\hat{p} \geq k/n   \pi)$	iscamonepropztest(k, n, pi, "greater") iscamonepropztest(k, n, pi, "less")
Two-tailed probability using Normal z-test, $P(\hat{p} \leq k/n \text{ or } \hat{p} \geq (1-k)/n   \pi)$	iscamonepropztest(k, n, pi, "two.sided")
95% (two-sided) Normal confidence interval	iscamonepropztest(k, n, conf.level = c(90, 95))

\* In addition to numerical output these functions provide a graphical representation.

## Chapter 1 Appendix - Stratified Random Sampling

Another way to reduce variability without taking larger samples is to take even more care in our sampling. For example, a *stratified sampling method* splits the population into homogenous groups first, and then samples a preset proportion from each subgroup. In the Gettysburg Address example, if we suspect nouns tend to be longer than non-nouns but worry that with only 16% nouns in the population we could easily end up with a sample without nouns, we can force the sample to contain 3 nouns and 17 non-nouns. This method will again be unbiased and if we stratify on a useful variable, we should find even less random sampling variability. Below we see in this case that in stratified random samples of size 20, there is a little bit less variability of sample proportions (though not much here).



## CHAPTER 2: ANALYZING QUANTITATIVE DATA

This chapter parallels the previous one in many ways. The difference here is that these investigations will involve a *quantitative* variable rather than a *categorical* one. This requires us to learn different tools for graphing and summarizing our data, as well as for statistical inference. In the end, you will find that the basic concepts and principles that you learned in Chapters 1 still apply.

### Section 1: Descriptive Statistics

Investigation 2.1: Birth weights – Normal model, Assessing model fit

Investigation 2.2: How long can you stand it? – Skewed data

Investigation 2.3: Cancer pamphlets - Application

### Section 2: Inference for Mean

Investigation 2.4: The *Ethan Allen* – Sampling distributions for  $\bar{x}$

Investigation 2.5: Healthy body temperatures – One-sample *t*-procedures

Probability Detour: Student's *t* Distribution

Investigation 2.6: Healthy body temperatures (cont.) – Prediction intervals

### Section 3: Inference for Other Statistics (optional)

Investigation 2.7: Water oxygen levels – Sign test

Investigation 2.8: Turbidity – *t*-procedures with transformed data

Investigation 2.9: Heroin treatment times - Bootstrapping

Example 2.1: Pushing On – One-sample *t*-procedures

Example 2.2: Distracted Driving? – Sign test

## SECTION 1: DESCRIPTIVE STATISTICS

In this section, we will begin our exploration of numerical and graphical summaries when the variable is quantitative. You will learn a new set of tools, but keep in mind that the overall goal is to informatively summarize the data and to let the data tell their story.

### Investigation 2.1: Birth Weights

The CDC's [Vital Statistics Data](#) allows you to download birth records for all births in the U.S. in a particular year. In fact, we downloaded the records for all 3,940,764 births in 2013 and then extracted several variables including the birth weight of the child (in grams). Can we use these data to build a *model* of how birth weights can be expected to behave in the future? Can we use that model to make predictions about certain kinds of birth weights? The file [USbirthsJan2013.txt](#) contains information on all the births in January 2013, including birth weight, whether the baby was full term (gestation over 36 weeks), the 5 minute [apgar score](#) (an immediate measure of the infant's health), and the amount of weight gained by the mother during pregnancy (in lbs).

(a) Are these data likely to be representative of birth weights for all 3,940,764 U.S. births in 2013? Explain.

(b) Is the variable *birthweight* quantitative or categorical?

Our next step is to look at the data! As in Chapter 1, we will want to consider which graphical and numerical summaries reveal the most information about the distribution.

**Accessing quantitative data:** See the Technology Detour after Investigation 1.1 for instructions on how to load data into R/Minitab. Remember to search the help menus for more information if you have more complicated data files in the future.

**R Notes:** Recall that the + sign comes up in R after you hit Enter when R is expecting the next line to be a continuation of the previous line. You can alternatively continue the command on the same line. In the instructions below, some steps are optional, but we wanted you to know they were there. *You should get into the habit of “attaching” the data once it has been read into R.*

Check that you have 324,314 rows of data.

Now use technology to create a *dotplot* of the birth weights.

### Technology Detour – Dotplots

#### In R Assuming you imported the data into “births”

```
> attach(births); names(births)      ← Shows the variable names for your data
> iscamdotplot(birthweight)          ← Input quantitative variable (after attaching)
```

(c) Describe what you see.

**Discussion:** The observations at 9999 don't seem to belong. The “[codebook](#)” for these data states that the largest birth weight is 8165 grams and for other variables it lists 99, 999, 9999 as values for “not stated” or unknown.

463-466	4	DBWT	Birth Weight – Detail in Grams	U,R	0227-8165 Number of grams
25-28	4	DOB_TT	Birth Time	R	0000-2359 Time of Birth 9999 Not Stated

You could convert these observations to the missing value designator in your software (e.g., \* in Minitab and NaN in R) or you can create a new data set that does not include those rows.

### Technology Detour – Subsetting the Data

#### In R

```
> births2=births [which(birthweight < 9999), ]
   ← selects rows that satisfy condition
   ← notice the comma!
> attach(births2)
   ← “birthweight” now refers to these data
Or, you can refer to the new data using births2$birthweight
```

(d) Recreate the dotplot with the subsetted data and describe what you see.

It may still be difficult to see much in the dotplot with such a large data set, especially if there are many distinct (non-repeated) values. One solution is to “bin” the observations. Some software packages (e.g., Minitab) will do that automatically even with a dotplot. Another approach is to use a different type of graph, a [histogram](#), that groups the data into intervals of equal width (e.g., 1000-2000, 2000-3000, ...) and then construct a bar for each interval with the height of the bar representing the number or proportion of observations in that interval. Notice these bars will be touching, unlike in a bar graph, to represent the continuous rather than categorical nature of the data.

## Technology Detour – Histograms

### In R

- Choose **Packages** in the lower right box menu bar and then check the box for the **Lattice** package.
- Then in the R Console window, type  
`> histogram(birthweight)`

**Note:** You may first have to install the package.

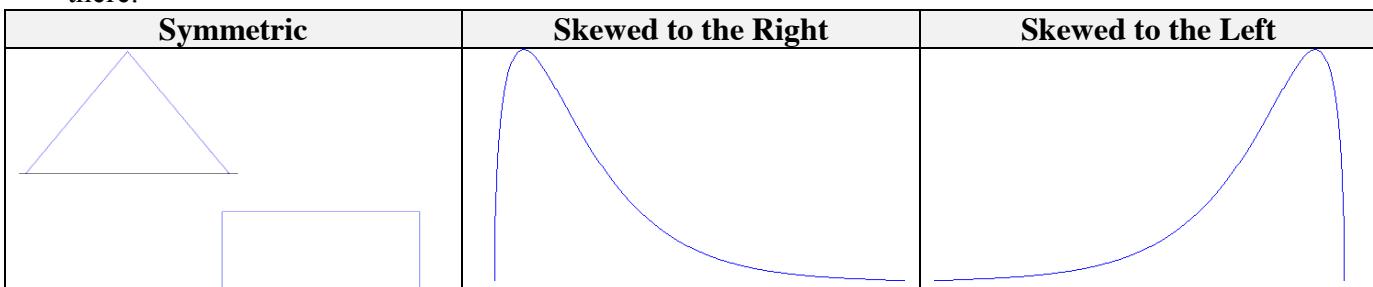
Or > `library("lattice")` once it's installed.

Notice the default is to give you the percentage of observations in each bin

- (e) Create a histogram for the subsetted data and compare the information revealed by the dotplots and the histograms. Do you feel one display is more effective at displaying information about birth weights than the other? Explain.

**Discussion:** In describing the *distribution* of quantitative data, there are three main features:

- Shape:* What is the pattern to the distribution? Is it symmetric or skewed? Is it unimodal or is there more than one main peak/cluster? Are there any individual observations that stand out/don't follow the overall pattern? If so, investigate these “outliers” and see whether you can explain why they are there.



- Center:* Where is the distribution centered or clustered? What are typical values?
- Variability:* How spread out is the distribution? How far do the observations tend to fall from the middle of the distribution? What is the overall range ( $\max - \min$ ) of the distribution?

It is also very important to note any observations that do not follow the overall pattern (e.g., “outliers”) as they may reveal errors in the data or other important features to pay attention to.

- (f) How would you characterize the shape, center, and variability of these birthweights? (Remember to put your comments in context, e.g., “The distribution of January birthweights in 2013 was ...”)

- (g) You may have noticed a few more lower weight babies than we might have expected (if we assume a “random” biological characteristic will be pretty symmetric and bell-shaped). Can you suggest an explanation for the excess of lower birth weights seen in this distribution?

(h) Now, further subset the births2 data based on whether or not the pregnancy lasted at least 37 weeks (i.e., full\_term = 1). *Note:* full\_term values of 2 indicate missing values for this variable.

[*R Hints:* use “==” for an equality comparison or be clever in your inequality. Attach and use length(birthweight) to count observations.]

How many observations do you end up with?

(i) After this step do we have a more symmetric distribution? What is a downside to subsetting the dataset in this way?

Let's use this as our final dataset and use technology to calculate some helpful numerical summaries of the distribution, namely describing the center and variability.

### Technology Detour – Numerical Summaries

#### In R

```
> attach(births3)                                Or use births3$birthweight
> iscamsummary(birthweight, digits=4)
Entering "digits =" specifies the number of significant digits you want displayed. Default is 3.
```

(j) Calculate and interpret the mean and the standard deviation of these data.

(k) How would the mean and standard deviation compare if we had not removed the 9999 values?

Recall from Chapter 1 the normal probability distribution has some very nice properties and allows us to predict the behavior of our variable. For example, we might want to assume birth weights follow a normal distribution and estimate how often a baby will be of low birth weight (under 2500 grams according to the *International Statistical Classification of Diseases, 10<sup>th</sup> revision*, World Health Organization, 2011) based on these data.

(l) Do these birth weight data appear to behave like a normal distribution? How are you deciding?

### Assessing Model Fit

Many software programs will allow you to overlay a theoretical normal distribution to see how well it matches your data, using the mean and standard deviation from the observed data.

(m) Overlay a normal model on the distribution of birthweight data and comment on how well the model fits the data:

- **In R**

> iscamaddnorm(birthweight) **Will take a few seconds to appear**

Discuss any deviations from the pattern of a normal distribution.

Another method for comparing your data to a theoretical normal distribution is to see whether certain properties of a normal distribution hold true. In Chapter 1, you learned that 95% of observations in a normal distribution should fall within two standard deviations of the mean.

(n) Calculate the percentage of the birthweights that fall within 2 standard deviations of the mean by creating a *Boolean* (true/false) variable:

- **In R**

```
> within2sd = (birthweight > mean(birthweight) -  
  2*sd(birthweight)) & (birthweight < mean(birthweight) +  
  2*sd(birthweight))  
> table(within2sd)
```

(o) How well does this percentage match to what would be predicted if the data were behaving like a normal distribution?

Another way to assess the “fit” of a probability model to the data is with a “probability plot” (aka quantile plot). Probability plots work roughly like this: The computer determines how many observations are in the data set, and then reports the  $z$ -scores it would expect for the  $1/n^{\text{th}}$  percentiles for (also called *quantiles*) for the probability model. The plot compares these  $z$ -scores to the observed data; if they “line up” then the graph supports that the data are behaving like a normal distribution. Whether or not the observations follow a line can be easier to judge than whether a histogram follows a curve.

(p) Use technology to create a normal probability plot for the subsetted birth weight data.

- **In R**

> qqnorm(birthweight, datax=T) **“datax” puts observed values on x-axis**

Do the observations deviate much from a line? If so how? [Hints: What does this suggest about *how* the birth weight data values differ from what we would predict for a normal distribution (e.g., smallest values are smaller than expected or larger than expected)? Is this consistent with what you saw in the histogram?]

**Discussion:** Some software packages report a p-value with the normal probability plot. The null hypothesis here is actually that the data *do* follow a normal distribution, so if you fail to reject this null hypothesis you can say that the data do not provide strong evidence that they do not arise from a normally distributed population. (But keep in mind that large sample sizes will drive the p-value down, regardless of the actual shape of the population.) There are several different types of significance tests for normality, but in this text we will focus on the visual judgement of whether the probability plot appears to follow a straight line.

Suppose we were willing to assume that in general birth weights are approximately normally distributed with mean 3330 grams and standard deviation 475 grams. We can use this *model* to make predictions, like how often a baby will be of *low birth weight*, defined as 2500 grams or less.

(q) Use technology (e.g., Technology Detour with Investigation 1.8) to calculate the normal distribution probability that a randomly selected baby will be of low birth weight. Include a well-labeled sketch with the shaded area of interest.

(r) Now examine the births3 data: What percentage of the birth weights in this data set were at most 2500 grams? How does this compare to the prediction in (q)? Does this surprise you? Explain. [Hint: Create a Boolean variable?]

**Discussion:** We can often use a theoretical model to predict how data will behave. However, it is often unclear whether the model we are using is appropriate. Sometimes we have to consider the context (e.g., biological characteristic) and presume a particular model. We can also use existing data to create a model but we need to consider what population the data are representative of and how stable we think the data generating process is (e.g., not changing over time). In this case, a normal distribution appears to underestimate the likelihood of smaller and heavier babies, so it could be very risky to make predictions like we did in (r).

**Practice Problem 2.1A**

Millions of people from around the world flock to Yellowstone Park in order to watch eruptions of Old Faithful geyser. But how long does a person typically have to wait between eruptions? How predictable is the geyser? The file [OldFaithful2011.txt](#) contains data on all eruptions in 2011 (from Electronic Monitor Files at <http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFATHFUL>). The *intereruption* column is the amount of time that passed between eruptions.

- (a) Create a histogram of the *times between eruptions*. Also calculate the mean and standard deviation of these intereruption times.
- (b) Does the mean appear to be a reasonable summary of how long you might expect to wait on average? What might you report instead?
- (c) Would you rather visit a geyser with a low mean and a large standard deviation, or a high mean and a low standard deviation? Explain.
- (d) Researchers have found that the length of the previous eruption (e.g., more than 3 minutes long or less than 3 minutes long) appears to be a good predictor of the time until the next eruption. Explain how this graph might support that statement.
- (e) Suppose we consider the times between 75 and 125 minutes. Do they appear to follow a normal distribution?

**Practice Problem 2.1B**

An expert witness in a paternity suit testifies that the length (in days) of pregnancy (the time from conception to delivery of the child) is approximately normally distributed with mean  $\mu = 270$  days and standard deviation  $\sigma = 10$  days. The defendant in the suit is able to prove that he was out of the country during a period that began 280 days before the birth of the child and ended 230 days before the birth of the child.

- (a) Does a normal model seem to be a reasonable assumption here? Explain why or why not or how you might decide.
- (b) If the defendant was the father or the child, what is the probability that the mother could have had the very long (more than 280 days) or the very short (less than 230 days) pregnancy indicated by the testimony?

## Investigation 2.2: How Long Can You Stand It?

Diekmann, Krassnig, and Lorenz (1996) conducted a field study to explore whether driver characteristics are related to an aggressive response (Thanks to Jeff Sklar for pointing us to this article). The study was conducted at a busy intersection in Munich, West Germany, on two afternoons (Sunday and Monday) in 1986. The experimenters sat in a Volkswaggen Jetta (the “blocking car”) and did not accelerate after the traffic light turned green, and timed how long before the blocked car driver reacted (either by honking or flashing headlights). The response time (in seconds) is our variable of interest. Some values were “censored” in that the researcher stopped timing before the driver actually honked. This can happen if there is a time limit to the observation period and “success” has not been observed within that time period.

- (a) How long do you think you would wait before you honked?

(b) The data can be found in [honking.txt](#). Use technology to create a histogram and describe the behavior of the data – shape, center, spread, outliers (suggest an explanation?). Then overlay a normal probability model. Do these data behave like a normal distribution? If not, how do they deviate from normality? Also examine a normal probability plot and discuss how deviations from the line correspond to the normal name shape you are observing. [*Hint:* Were the observed response times/quantiles placed on the vertical or the horizontal axis?]

(c) Compare the mean and median weight times. Which is larger? Why is it larger?

**Definition:** A numerical summary (statistic) is said to be resistant when it is not strongly influenced by a change in one or two extreme data values.

When data are skewed to the right, the mean will be pulled in the direction of the larger values.

When data are skewed, we might often prefer to report the median as a “typical” value in the data set, rather than the mean which is pulled in the direction of the longer tail. In addition, you might not want to cite the standard deviation as a “symmetric” measure of spread in the distribution.

**Definition:** Interquartile range (IQR) = upper quartile – lower quartile

The lower quartile is a value such that roughly one-fourth of all the observations fall below it; the upper quartile is a value such that roughly one-fourth of all the observations fall above it. The IQR then measures the width of the interval containing the middle 50% of the observations.

- (d) Use technology to compute the interquartile range (*Hint*: Run the descriptive statistics command and report the lower and upper quartiles and then subtract). Write a one-sentence interpretation of this value. Would you consider it a resistant measure of spread? Explain.

When the data are skewed, the median and interquartile range are often considered better numerical summaries of the center and variability of the distribution. When working with the median and interquartile range, we often report the [five number summary](#) which consists of the minimum, lower quartile, median, upper quartile, and maximum values.

**Definition:** Another graph is based on the five-number summary, called a [boxplot](#) (invented by John Tukey in 1970). The box extends from the lower quartile to the upper quartile with a vertical line inside the box at the location of the median. Whiskers then typically extend to the min and max values.

- (e) Create by hand a boxplot for these data. Which display do you prefer, the boxplot or the histogram? Why?

Although boxplots are a nice visual of the five-number summary, they can sometimes miss interesting features in a data set. In particular, shape can be more difficult to judge in a boxplot.

Another application of the inter-quartile range is as a way to measure whether an observation is far from the bulk of the distribution.

**Definition:** A value is an [outlier](#) according to the *1.5IQR criterion* if the value is larger than the upper quartile  $+ 1.5 \times$  box length or smaller than the lower quartile  $- 1.5 \times$  box length.

Note: The *box length* = upper quartile – lower quartile, is called the [interquartile range](#) or IQR.

A [modified boxplot](#) will display such outliers separately and then extend the whiskers to the most extreme non-outlier observation.

- (f) Use the following Technology Detour to create a “modified” boxplot for these data. Are there any outliers?

### Technology Detour – Modified Boxplots

#### In R

```
> boxplot(responsetime, ylab="time until reaction"           ← Adds labels
+ horizontal=TRUE)                                         ← Makes horizontal
OR
> iscamboxplot(responsetime)                                ← Uses quartiles
```

## Modelling non-normal data

These data are not well modelled by a normal distribution. So can we still make predictions? There are a couple of strategies. One would be to consider whether a *rescaling* or *transformation* of the data might create a more normal-looking distribution, allowing us to use the methods from Investigation 2.1. In this case, we need a transformation that will downsize the large values more than the small values. Log transformations are often very helpful in this regard.

**Definition:** A [data transformation](#) applies a mathematical function to each value to re-express the data on an alternative scale. Data transformations can also make the data more closely modeled with a normal distribution, which could then satisfy the conditions the Central Limit Theorem and inference procedures based on the *t*-distribution.

(g) Create a new variable which is  $\log(\text{responsetime})$ . (You can use either natural log or log base 10, but so we all do the same thing, let's use natural log here, which is the default in most software when you say "log.")

- **In R**

```
> lnresponsetime = log(responsetime)
```

Create a histogram of these data and a normal probability plot. Does  $\log(\text{responsetime})$  approximately follow a normal distribution? What are the mean and standard deviation of this distribution?

(h) Use a normal distribution with mean 1.29 ln-sec and standard deviation 0.53 ln-seconds for the *logged response times* and predict how often someone will honk within the first 2 seconds. (*Hint:* What are you going to use for the *observation* of interest?)

(i) How does this prediction in (h) compare to the observed percentage honking within the first 2 seconds in the data set?

Another approach is to fit a different mathematical model to the original data:

(j) Use technology to overlay an *exponential* probability model (often used to model wait times) to these data and/or create a probability plot using the exponential distribution as the reference distribution.

- **In R** (for the qqplot we have to first get the quantiles)

> theoquant = qexp(ppoints(12))	← Generates 1/n quantiles for 12 observations from exponential distribution
> histogram(theoquant)	
> qqplot(responsetime, theoquant)	← Your data vs. quantiles. Look for a line.
> iscamaddexp(responsetime)	← overlay exponential model

Describe the behavior of the exponential distribution. Does it appear to be a reasonable fit for these data? Describe any deviations.

(k) Use technology to calculate the probability of a wait time under 2 seconds using the exponential distribution with mean 4.25 sec

- **In R**

```
> pexp(2, rate = 1/4.25)           ← P(X ≤ 2) when X ~ Exp(mean = 4.25)
```

How does it compare to your estimate in (h) and the result in (i)?

(l) Repeat (j) and (k) with a “lognormal” distribution.

- **In R**

```
> qqplot(responsetime, qlnorm(ppoints(12)))
> iscamddlnorm(responsetime)
> plnorm(2, meanlog = 3.383, sdlog = 0.5285)
```

Note: This is equivalent to fitting the normal distribution on the log-transformed data!

**Discussion:** There are of course, many other probability models we could look into. One limitation of the exponential distribution is assuming the mean and standard deviation are equal, clearly not the case for these data. There are other more flexible distributions (e.g., Gamma and Weibull) that use two parameters to characterize the distribution rather than only one.

(m) To what population are you willing to generalize these results? Explain. [Hint: Think about how the data were collected and what you learned about sampling methods in Chapter 1.]

(n) If you were to continue to explore this research area/ data set, what would you be interested in investigating next?

## Study Conclusions

In this study, we found that the amount of time a blocked driver waits before responding follows a skewed right distribution with a mean of 4.25 seconds and a median of 3.24 seconds, with a few drivers waiting more than 10 seconds. Although the dataset is small, we might consider using these data to build a mathematical model (for these skewed data either using a transformation or a probability model other than the normal distribution) to help predict future results (e.g., a wait time of less than 2 seconds). These researchers were actually interested in whether the “social status” of the blocked car was related to how long it took before people honked. They found that “the mean and median response times decreased monotonically with the status of car, except when the blocked car was very small.” Similar results had been found in an earlier study by Doob and Gross in the United States (1968) which varied the status of the blocking car. However, neither study was replicated by a Swiss study (Jann, Suhner, & Marioni, 1995), perhaps due to cultural differences impacting generalizability.

## Practice Problem 2.2A

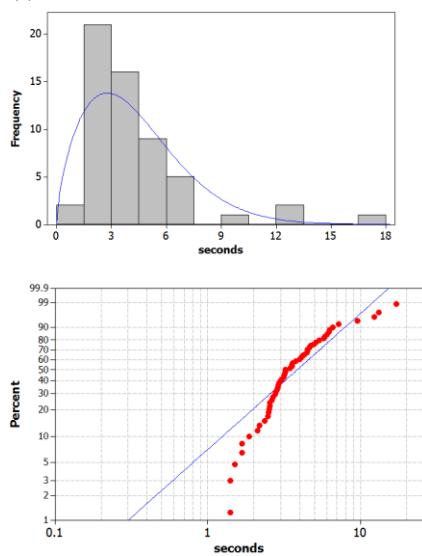
A group of Cal Poly students (Sasscer, Mease, Tanenbaum, and Hansen, 2009) conducted the following study: volunteers were to say “go,” and then to say “stop” when they believed 30 seconds had passed. The researchers asked the participants not to count in their heads and recorded how much time had actually passed. The data are in [30seconds.txt](#).

- Examine graphical and numerical summaries of these data. Describe the shape of the distribution. How do the mean and median compare? What does this tell you about whether people tend to over or underestimate the length of 30 seconds?
- Does a normal model seem appropriate here?
- Does a log transformation succeed in creating a normal distribution? Explain how you might have predicted this answer based on the first graph you looked at.
- To what population would you be willing to generalize these data?

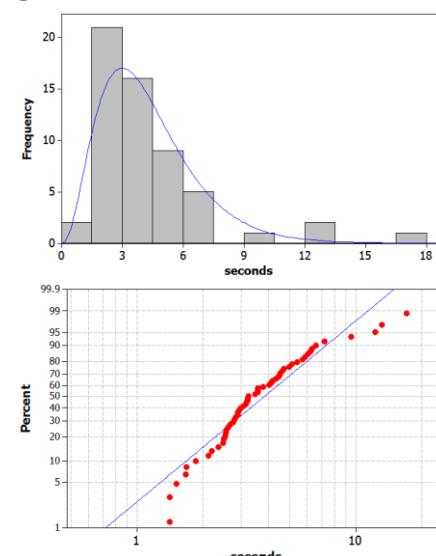
## Practice Problem 2.2B

Below are distribution fits for the honking data with a Weibull distribution and a Gamma distribution.

### Weibull



### Gamma



Which do you think indicates a better fit for these data? Explain how you are deciding.

### Investigation 2.3: Readability of Cancer Pamphlets

Researchers in Philadelphia investigated whether or not pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend. They applied tests to measure the reading levels of 63 cancer patients and also the readability levels of 30 cancer pamphlets (based on such factors as sentence length and number of polysyllabic words). These numbers correspond to grade levels, but cancer patient reading levels below grade 3 and above grade 12 were not determined exactly.

The following tables indicate the number of patients at each reading level and the number of pamphlets at each readability level:

Patient's reading level	<3	3	4	5	6	7	8	9	10	11	12	>12
Count	6	4	4	3	3	2	6	5	4	7	2	17

Pamphlet's readability level	6	7	8	9	10	11	12	13	14	15	16
Count	3	3	8	4	1	1	4	2	1	2	1

- (a) How many of these patients had a 6<sup>th</sup> grade reading level?
- (b) Explain why the form of the data does not allow you to calculate the mean reading skill of these cancer patients.
- (c) Determine median reading level for these patients. (*Hint:* Consider the counts, and remember there are 63 patients.)
- (d) Determine the median readability level of these pamphlets.
- (e) How do these medians compare? Are the values fairly close?
- (f) Does the closeness of these medians indicate the pamphlets are well matched to the patients' reading levels? Explain. (*Hint:* You may want to perform some additional explorations of these data?) How would you respond to these researchers?

**Discussion:** Keep in mind when examining quantitative data, that you should start by constructing some simple graphs to explore the data. In fact, you should use technology to explore a couple of different graphs (e.g., change the bin width in a histogram) to help reveal hidden patterns and unusual observations.

### Study Conclusions

When these researchers came to the statistical consultant, they wanted a p-value for comparing the mean reading level of the patients to the mean reading level of the pamphlets. You will learn about such a test in Chapter 4. However, the statistical consultant replied that not only couldn't means be calculated, but that a simple look at the data revealed a substantial percentage of the patients with a reading level below that of the simplest pamphlet. This "descriptive" analysis was sufficient, rather than looking for a more complicated "inferential" analysis involving p-values and confidence intervals.

### Practice Problem 2.3

- If you want to estimate total tax revenue for a city (how much money people will pay in based on their incomes), would you rather know the mean income or the median income? Explain.
- If you want to estimate the typical housing cost for a new city, would you rather know the mean housing cost or the median housing cost? Explain.
- In the Cancer Pamphlet study, explain why a comparison of the centers of the distributions does not match the research question of interest.

## SECTION 2: INFERENCE FOR POPULATION MEAN

In the previous section, you transitioned from categorical data to quantitative data. We started by looking at graphs and numerical characteristics of the distribution, and here, we found that looking at the variability of the distribution was often quite an important feature. With proportions, the sample proportion tells us everything about the shape, center, and variability of the sample distribution. But with quantitative data we will need to consider all three of these characteristics, as well as possible outliers and other unusual observations (e.g., clustering). But often, like in Chapter 1, we want to make inferences beyond our sample data to a larger population or process. In particular, with the Gettysburg Address data, we saw that there is also a very predictable pattern to how sample means vary from random sample to random sample, allowing us to estimate how close we expect our observed statistic (e.g., the sample mean) to be to the unknown population parameter (e.g., the population mean). We will now further explore these theoretical results before considering other types of parameters as well (e.g., the population median). The reasoning will be very similar but we will need to consider a new probability model to take into account that we will also be estimating the variability in our quantitative variable.

### Investigation 2.4: The *Ethan Allen*

On October 5, 2005, a tour boat named the Ethan Allen capsized on Lake George in New York with 47 passengers aboard. In the inquiries that followed, it was suggested that the tour operators should have realized that the combined weight of so many passengers was likely to exceed the weight capacity of the boat. Could they have predicted this?

- (a) The variable of interest in this study is *weight of a passenger*. For a sample of 47 passengers, make a conjecture as to the distribution (shape, center, variability) of this variable. Include a well-labeled sketch of your conjecture below (e.g., 47 dots or sketch an outline of the general pattern). Include a guess of the values of the sample mean  $\bar{x}$  and the sample standard deviation  $s$ . [Hint: Think about the size of a typical deviation from the mean, as well as the Empirical (95%) Rule.]

- (b) Data from the Centers for Disease Control and Prevention indicate that weights of American adults in 2005 had a mean of 167 pounds and a standard deviation of 35 pounds. (To convey that these are population values, we will use Greek letters to represent their values,  $\mu = 167$  and  $\sigma = 35$ .) Use this information to sketch a possible distribution of the weights of the population of adult Americans.

The maximum weight capacity of passengers that the Ethan Allen could accommodate was estimated to be 7500 pounds. If the tour boat company consistently accepted 47 passengers, what we want to know is the probability that the combined weight of the 47 passengers would exceed this capacity.

(c) What are the observational units and variable for this research question? Is this variable quantitative or categorical? [Hint: Notice that the probability of interest does not pertain to individual passengers.]

(d) Instead of focusing on the *total* weight of the 47 passengers, we can instead work with the *average* weight of the 47 passengers. Complete the following to indicate how these are equivalent:

*Total weight of passengers* > 7500 lbs      when      *average weight of 47 passengers* > \_\_\_\_\_.

So, to see how often the boat was sent out with too much weight, we need to know about the distribution of the *average weight of 47 passengers* from different boats (samples). Think about a distribution of sample mean weights from different random samples of 47 passengers, selected from the population of adult Americans.

(e) Where do you think the distribution of sample means will be centered?

(f) Do you think the distribution of sample means would have more variability, less variability, or the same variability as weights of individual people? Explain your answer.

(g) Do you think the probability of having the *average* weight exceed 159.574 pounds is larger or smaller than the probability of the weight of any *one* passenger exceeding 159.574 pounds?

To investigate this probability, we will generate random samples from a hypothetical population of adults' weights. Open the [WeightPopulations.xls](#) data file and pretend each column is a different population of 20,000 adult tourists, each with a population mean weight ( $\mu$ ) of 167 lbs and a population standard deviation ( $\sigma$ ) of 35 lbs.

- Copy the data in the first column (`pop1`) to the clipboard and then open the [Sampling from Finite Population](#) applet: Press **Clear**, Click inside the **Paste population data** box, and paste the data from the clipboard. Press **Use Data**.

(h) Describe the shape, mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of this population distribution (shown the histogram below the window).

Paste population data c

ID	pop1
1	187
2	180
3	100
4	197
5	175
6	169
7	168
8	170
9	149

Use Data    Clear    To

Population size: 20000

- Check the **Show Sampling Options** box. Keep the **Number of Samples** at 1 and specify the **Sample Size** to be 47 and press the **Draw Samples** button.

Show Sampling Options:

Number of Samples:

Sample Size:

**Draw Samples** **Reset**

The distribution of weights for this random sample of 47 passengers from this population is displayed in blue and in the *Most Recent Sample* graph.

- (i) Describe the shape, center, and variability of this distribution. How does this sample distribution compare to the population distribution?

Shape:

Mean:

Standard deviation:

Comparison (samples vs. population):

Notice that the mean of this sample ( $\bar{x}$ ) has been added to the graph in the lower right *Statistic* graph.

- (j) Press **Draw Samples** again. Did you obtain the same sample of 47 weights? Did you obtain the same sample mean? Do either of the two sample means generated thus far equal the population mean?

- (k) Continue to press **Draw Samples** 8 more times (for 10 samples total), and notice the variability in the resulting sample means from sample to sample. Is a pattern beginning to emerge in the *Statistic* distribution graph?

- (l) Now set the **Number of Samples** to 990 (for 1000 total) and press **Draw Samples**. Describe the shape, center, and variability of the distribution of the sample means (lower right). How do these compare to the population distribution? [Hints: Note the change in scaling. You can check the **Population scale** box to rescale back to the original population scaling and add the location of the population mean.] Which one of these features has changed the most, and how has it changed?

Shape:

Mean:

Standard deviation:

Comparison (sample means vs. population):

- (m) How can we use your simulated distribution of sample means to decide whether it is surprising that a boat with 47 passengers would exceed the (average) weight limit by chance (random sampling error) alone?

(n) To investigate the question posed in (m), specify the sample mean of interest (159.574) in the **Count Sample** box, use the pull-down menu to specify whether you want to count samples *Greater Than*, *Less Than*, or *Beyond* (both directions). Press **Count**. What conclusions can you draw from this count?

## Non-normal Population

To carry out the preceding simulation analysis, we assumed that the population distribution had a normal shape. But what if the population of adult weights has a different, non-normal distribution? Will that change our findings?

(o) Now copy and paste the data from the “pop2” column in the WeightPopulations.xls data file. Describe the shape of this population and what it means for the variable to have this shape in this context. How do the values of  $\mu$  and  $\sigma$  compare to the previous population?

Shape: Mean: Standard deviation:

Mean:

Standard deviation:

### Comparison:

(p) Generate 1000 random samples of 47 individuals from this population. Produce a well-labeled sketch of the distribution of sample means below and note the values for the mean and standard deviation. How does this distribution compare to the one in (I)?

Sketch:  Mean:

SD.

Con

(q) Again use the applet to approximate the probability of obtaining a sample mean weight of at least 159.574 lbs for a random sample of 47 passengers from this population. Has this probability changed considerably with the change in the shape of the distribution of the population of weights?

(r) Repeat this analysis using the other population distribution ([pop3](#)) in the data file and summarize your observations for the three populations in the table below.

Distribution of sample means	Shape	Mean	Standard deviation
Normal ( $\mu = 167$ , $\sigma = 35$ )			
Skewed right ( $\mu = 167$ , $\sigma = 35$ )			
Uniform ( $a = 106.4$ , $b = 227.6$ )			

(s) Now consider changing the sample size from 47 to 188 (four times larger). Make a prediction for how the shape, mean, and SD of the distribution of sample means would change (if at all).

(t) Make this change of sample size from 47 to 188, and generate 1000 random samples (from the uniformly distributed population, `pop3`). Did the shape of the sample means change very much? What about the mean of the sample means? What about the SD of the sample means? How were your predictions in (s)?

**Discussion:** You should see that, as with the Gettysburg Address investigation, the *shape* of the population is not having much effect on the distribution of the sample means! In fact, you can show that the mean of the distribution of sample means from random samples is always equal to the mean of the population (any discrepancies you find are from not simulating enough random samples) and that, assuming the population is large compared to the size of the sample, the standard deviation of the distribution of sample means is equal to  $\sigma / \sqrt{n}$ . This standard deviation formula applies when the population size is large (more than 20 times the size of the sample) or infinite (so the randomly selected observations can be considered independent).

(u) Explain why the formula for the standard deviation of the sample mean ( $\sigma / \sqrt{n}$ ) makes intuitive sense (both the  $\sigma$  component and the  $\sqrt{n}$  component).

**Key Result:** If all possible samples of size  $n$  are selected from a large population or an infinite process with mean  $\mu$  and standard deviation  $\sigma$ , then the *sampling distribution of these sample means* will have the following characteristics:

- The mean will be equal to  $\mu$ .
- The standard deviation will be equal to  $\sigma / \sqrt{n}$  (we can call this  $SD(\bar{x})$  or  $\sigma_{\bar{x}}$ ).
- **Central Limit Theorem:** The shape will be normal if the population distribution is normal, or approximately normal if the sample size is large regardless of the shape of the population distribution.

The convention is to consider the sample size large enough if  $n \geq 30$ . However, this rule really depends on the shape of the population. The more non-symmetric the population distribution, the larger the sample size necessary before the distribution of sample means is reasonably modeled by a normal distribution.

The population is considered large if it is more than 20 times the size of the sample.

**Discussion:** In general, the *shape* of the distribution of sample means does *not* depend on the shape of the population distribution, unless you have small sample sizes. So if the population distribution itself follows a normal distribution, then we will always be willing to model the distribution of sample means with a normal distribution. However, we typically don't know the distribution of the population (that's why we need to collect data), but we can make a judgment based on the nature of the variable (e.g., biological characteristics, repeated measurements) or based on the information conveyed to us by the shape of the distribution of the *sample* data (e.g., normal probability plot).

In this example, a sample size of 47 appears large enough to result in a normal distribution for the distribution of sample means. However, you may have noticed a bit of a right skew when the population was skewed and in such a situation you would want a larger sample size before you were willing to model the distribution of sample means with a normal distribution.

Keep in mind that the results about the mean and standard deviation always hold for random samples: sample means cluster around the population mean and are less variable than individual observations! If the population size is not large, than a “finite population correction factor” can be applied as in Ch. 1.

(v) Use the theoretical results for the mean and standard deviation of a *sample mean* to standardize the value of 159.57 lbs. [Hint: Start with a well-labeled sketch.] How might you interpret this value?

(w) Use the result from the Central Limit Theorem and technology (e.g, the [Normal Probability Calculator](#) applet) to estimate the probability of a sample mean weight exceeding 159.57 lbs for a random sample of 47 passengers from a population with mean  $\mu = 167$  lbs and standard deviation  $\sigma = 35$  lbs. [Hint: Shade the area of interest in your sketch in (u).] How does this estimated probability compare to what you found with repeated sampling from the hypothetical populations?

(x) Identify one concern you might have with this analysis. [Hint: What other assumption, apart from the shape of the population, was made in these simulations that may not be true in this study? Do you think this is a reasonable assumption for this study?]

**Study Conclusions**

Assuming the CDC values for the mean and standard deviation of adult Americans' weights,  $\mu = 167$  lbs and  $\sigma = 35$  lbs, we believe that the distribution of sample mean weights will be well modeled by a normal distribution (based both on the not extremely skewed nature of the variable and the moderately large sample size of 47, which is larger than 30). Therefore, the Central Limit Theorem allows us to predict that the distribution of  $\bar{x}$  is approximately normal with mean 167 lbs and standard deviation  $35/\sqrt{47} \approx 5.105$  lbs. From this information, assuming the CDC data is representative of the population of Ethan Allen travelers, we can estimate the probability of obtaining a sample mean of 159.57 pounds or higher to be 0.9264. Therefore, it is not at all surprising that a boat carrying 47 American adults capsized. In fact, the surprising part might be that it didn't happen sooner!

**Practice Problem 2.4A**

- Use the Sampling from Finite Population applet or the Central Limit Theorem to estimate the probability that the sample mean of 20 randomly selected passengers exceeds 159.57lbs, assuming a normal population with mean 167lbs and standard deviation 35lbs.
- Is the probability you found in (a) larger or smaller than the probability you found for 47 passengers? Explain why your answer makes intuitive sense.
- Repeat (a) assuming a uniformly distributed population of weights. How do these two probabilities compare? [Hint: Think about whether it is more appropriate to use the Sampling from Finite Population applet or the CLT to answer this question.]
- Explain why the calculation in (a) does not estimate the probability of the Ethan Allen sinking with 20 passengers.

**Practice Problem 2.4B**

Use the Sampling from Finite Population applet or the Central Limit Theorem to estimate the probability that the sample mean of 47 randomly selected passengers would exceed 159.57lbs, assuming that random samples are repeatedly selected from a population of 80,000 individuals with mean 167 lbs and standard deviation 35 lbs. State any assumptions you need to make, and support your answer statistically.

### Investigation 2.5: Healthy Body Temperatures

In the previous investigation, we assumed random sampling from a finite population to predict the distribution of sample means and help us evaluate whether a particular value is an unlikely value for the sample mean by chance alone. However, this method is not realistic in practice as we had to make up a population to sample from and we had to make some assumptions about that population (e.g., shape). Luckily, the Central Limit Theorem also predicts how that distribution would behave for most population shapes. But the CLT still requires us to know certain characteristics about the population.

What is a healthy body temperature? Researchers have cited problems with Carl Wunderlich's "axioms on clinical thermometry" and claimed that the traditional value of 98.6°F is out of date (Mackowiak, Wasserman, & Levine, *Journal of the American Medical Association*, 1992). Body temperatures (oral temperatures using a digital thermometer) were recorded for healthy men and women, aged 18-40 years, who were volunteers in Shigella vaccine trials at the University of Maryland Center for Vaccine Development, Baltimore. For these adults, the mean body temperature was found to be 98.249°F with a standard deviation of 0.733°F.

- (a) Explain (in words, in context) what is meant by the following symbols as applied to this study:  $n$ ,  $\bar{x}$ ,  $s$ ,  $\mu$ ,  $\sigma$ . If you know a value, report it. Otherwise, define the symbol.

$$n =$$

$$\bar{x} =$$

$$s =$$

$$\mu =$$

$$\sigma =$$

- (b) Write a null hypothesis and an alternative hypothesis for testing Wunderlich's axiom using appropriate symbols.

$$H_0:$$

$$H_a:$$

- (c) Suppose the axiom is correct and many different random samples of 13 adults are taken from a large normally distributed population with mean 98.6°F. What does the Central Limit Theorem tell you about the theoretical distribution of sample means? (Indicate any necessary information that is missing.)

If we assume the null hypothesis is true, then we have a value to use for the population mean. However, we don't have a value to use for the population standard deviation (sometimes called a "nuisance parameter" because we need its value to be able to use  $SD(\bar{x})$  but it's not the parameter of interest).

- (d) Suggest a method for estimating the population standard deviation from the sample data.

**Definition:** The standard error of the sample mean, denoted by  $SE(\bar{x})$ , is an estimate of the standard deviation of  $\bar{x}$  (the sample to sample variability in sample means from repeated random samples) calculated by substituting the sample standard deviation  $s$  for the population standard deviation  $\sigma$ :

$$SE(\bar{x}) = s / \sqrt{n}.$$

- (e) Calculate the value of the standard error of the sample mean body temperature for this study.
  - (f) Determine how many standard errors the sample mean (98.249) falls from the hypothesized value of 98.6.
  - (g) Based on this calculation, would you consider the value of the sample mean (98.249) to be surprising, if the population mean were really equal to 98.6? Explain how you are deciding.
- Previously, we compared our standardized statistics ( $z$ -scores) to the normal distribution and said (absolute) values larger than two were considered rare. Is that still true when we have used the sample standard deviation in our calculation? Let's explore the method you just used to standardize the sample mean (using the standard error) in more detail.
- (h) Open the [Sampling from a Finite Population](#) applet and paste the hypothetical population body temperature data from the [BodyTempPop.txt](#) file. Does this appear to be a normally distributed population? What are the values of the population mean and the population standard deviation?
  - (i) Use the applet to select 10,000 samples of **13** adults from this hypothetical population. Confirm that the behavior of the distribution of sample means is consistent with the Central Limit Theorem? [Hint: Discuss shape, center, and variability; compare predicted to simulated.]
  - (j) Where does the observed sample mean of 98.249 fall in this sampling distribution? Does it appear to be a surprising value if the population mean equals 98.6?

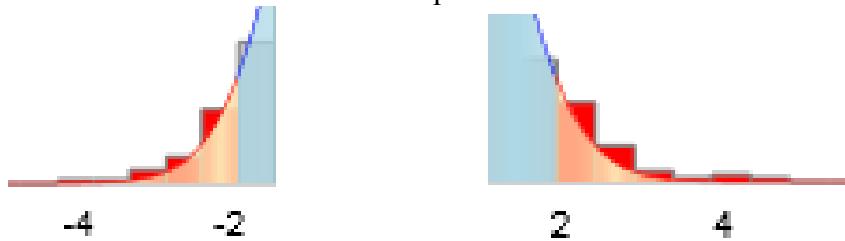
So if we have a value for  $\sigma$ , we are fine. But what about the statistic suggested in (f); does this standardized statistic behave nicely and is this distribution again well modeled by a normal distribution?

(k) In the applet, change the **Statistic** option (above the graph) to **t-statistic**, the name for the standardized sample mean using the standard error of the sample mean. Describe the shape of the distribution of these standardized statistics from your 10,000 random samples.

(l) Check the box to **Overlay Normal Distribution**; does this appear to be a reasonable fit? What p-value does this normal approximation produce? [Hint: Enter your answer to (f) as the observed result for the t-statistic and count beyond.]

(m) Does the theory-based p-value from the normal distribution accurately predict how often we would simulate a standardized statistic at least as extreme (in either direction) as the observed value of  $-1.73$ ? Does it over predict or underpredict? [Hint: How does the behavior of the distribution of the standardized statistics most differ from a normal model?]

**Discussion:** If we zoom in on the tails of the distribution, we see that more of the simulated distribution lies in those tails than the normal distribution would predict.



To model the sampling distribution of the standardized statistic  $(\bar{x} - \mu)/(s/\sqrt{n})$ , we need a density curve with *heavier tails* than the standard normal distribution. William S. Gosset, a chemist turned statistician, showed in 1908, while working for the Guinness Breweries in Dublin, that a “*t* probability curve” (see next page) provides a better model for the sampling distribution of this standardized statistic when the population of observations follows a normal distribution.

(n) Check the **Overlay t-distribution** box. What is the main visual difference in the *t*-distribution model compared to the normal distribution model? Does this *t*-distribution appear to be a better model for the simulated sampling distribution? Is the theory-based p-value using the *t*-distribution closer to the empirical p-value than the theory-based p-value using the normal distribution?

- (o) The actual body temperature study involved a sample of  $n = 130$  adults. Use the applet to generate a sampling distribution of  $t$ -statistics for this sample size. Toggle between the normal and  $t$  probability distributions. Do you see much difference between them? What is the actual value of the observed  $t$ -statistic? Where does it fall in this distribution? What do you conclude about the null hypothesis?

**Discussion:** The consequence of this exploration is that when we are estimating both the population mean and the population standard deviation, we will compare our standardized statistic for the sample mean to the  $t$ -distribution instead of the normal distribution for approximating p-values and confidence intervals. Although with larger sample sizes, the distinction will be quite minor.

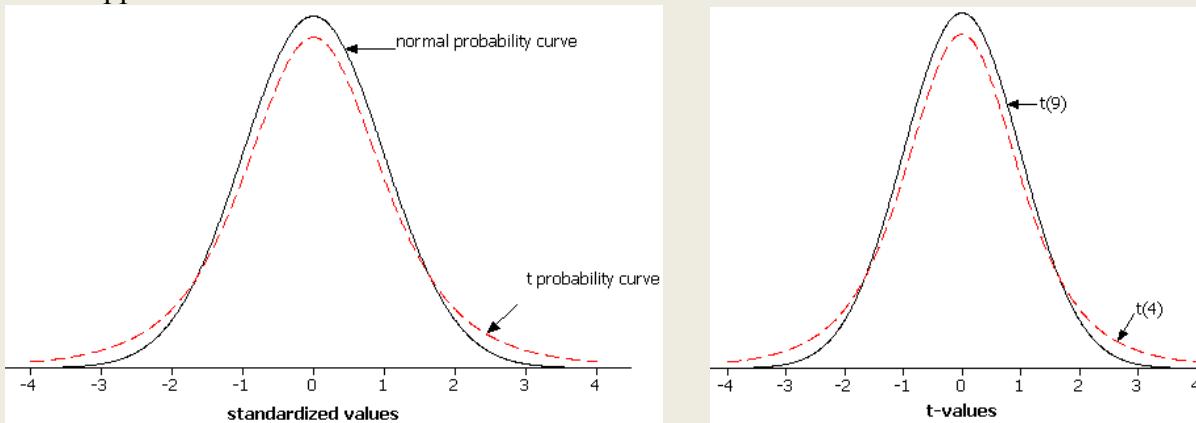
### Probability Detour – Student's $t$ -distribution

The  $t$  probability density curve is described by the following function:

$$f(x) \propto \frac{1}{\sqrt{v\pi}(1+x^2/v)^{(v+1)/2}} \text{ where } -\infty < x < \infty .$$

An impressive function indeed! But you should notice that this function only depends on the parameter  $v$ , referred to as the *degrees of freedom*.

This symmetric distribution has *heavier tails* than the standard normal distribution. We get a different  $t$ -distribution for each value of the degrees of freedom. As the degrees of freedom increase, the  $t$ -distribution approaches the standard normal distribution.



**One-sample  $t$ -test for  $\mu$ :** So to test a null hypothesis about a population mean when we don't know the population standard deviation (pretty much always),  $H_0: \mu = \mu_0$ , we will use the sample standard deviation to calculate the standard error  $SE(\bar{x})$  and compare the standardized statistic

$$t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

to a  $t$ -distribution with  $n - 1$  degrees of freedom. Theoretically, this approximation requires the population to follow a normal distribution. However, statisticians have found this approximation to also be reasonable for other population distributions whenever the sample size is large. How large the sample size needs to be depends on how skewed the population distribution is. Consequently, we will consider the  $t$  procedures valid when either the population distribution is symmetric or the sample size is large.

But what about confidence intervals?

(p) Turn to the [Simulating Confidence Intervals](#) applet. Change the **Method** to Means, but keep the population set to **Normal** and **z with  $\sigma$** . Set the population mean to 98.6, the population standard deviation to 0.733, and the sample size to 13. Generate 1000 random samples from this population and examine the running total for the percentage of 95% confidence intervals that successfully capture the actual value of the population mean  $\mu$ . Is this “z with sigma” procedure behaving as it should? How are you deciding?

(q) But more realistically, we don’t know  $\sigma$  and will use  $s$  in calculating our confidence interval. Predict what will change about the resulting confidence intervals from different random samples. [Hint: Think of two main properties of confidence intervals.]

(r) Change the **Method** now to **z with s**. What percentage of these 1000 confidence intervals succeed in capturing the population mean of 98.6? Is this close to 95%? If not, is it larger or smaller?

(s) Repeat (r) with a sample size of  $n = 5$ .

**Discussion:** This “z with  $s$ ” procedure produces a coverage rate (for successfully capturing the value of the population mean) that is less than 95%, because again the normal distribution doesn’t account for the additional uncertainty resulting from sample to sample when we need to use both  $\bar{x}$  and  $s$ . The fix will be to multiply the standard error by a critical value larger than 1.96 to compensate for the additional uncertainty introduced by estimating  $\sigma$  with  $s$ . The  $t$ -distribution will come to our rescue. Which  $t$ -distribution do we use? That will depend on our sample size; with smaller samples we need heavier tails and with larger samples we need a distribution more like the normal probability model. The heaviness of the tails will be determined by the “degrees of freedom” of the  $t$ -distribution.

(t) Change the **Method** to **t**. How do the intervals visibly change? Is the coverage rate indeed closer to 95%?

This leads to a confidence interval formula with the same general form as the previous chapter  $\text{sample statistic} \pm (\text{critical value}) \times (\text{SE of statistic})$ , where the critical value now comes from the  $t_{n-1}$  (degrees of freedom =  $n - 1$ ) distribution instead of the standard normal distribution.

**One-sample  $t$ -interval for  $\mu$ :** When we have a symmetric sample or a large sample size, an approximate confidence interval for  $\mu$  is given by:

$$\bar{x} \pm t_{n-1}^* s / \sqrt{n}$$

Keep in mind that the critical value  $t^*$  tells us how many standard errors we need to extend from the sample mean depending on how confident we want to be. Our goal is to develop a  $100 \times C\%$  confidence interval method that will capture the population parameter  $100 \times C\%$  of the time in the long run.

(u) Find the  $t^*$  value corresponding to a 95% confidence level and a sample size of  $n = 5$

- **In R:**  
`> iscaminvt(.95, 4, "between")`
- **t Probability Calculator applet:**
  - Specify 4 as the degrees of freedom
  - Check the box next to the less than symbol and then entire .025 in the probability box and press Return. The t-value box should fill in.

How does this critical value compare to the corresponding  $z^*$  value for 95% confidence?

$$t_4^* = \quad z^* =$$

Comparison:

(v) Repeat (u) for the sample size of  $n = 13$ . How do the  $t$ -critical values compare for these different sample sizes? Is this what you expected? Explain.

(w) Repeat (v) for the sample size of  $n = 130$ . How does this value compare to the earlier  $t^*$  and  $z^*$  values?

(x) Use the critical value from (w) to calculate a 95% confidence interval for the mean body temperature of a healthy adult based on our sample ( $\bar{x} = 98.249$ ,  $s = 0.733$ ). Is this interval consistent with your conclusion about the null hypothesis in (o)? Explain.

## Technology Detour – One Sample *t*-procedures

In R, you can use the `t.test` command with raw data ("x") or `iscamonesampel` with summary data

```
> t.test(x, mu = μ₀, alternative="two.sided", conf.level = .95)
```

OR

```
> iscamonesampel(xbar=98.249, sd=.733, n =130, hypothesized = 98.6, alternative = "two.sided", conf.level = 95)
where sd = sample standard deviation, s.
```

### In Theory-Based Inference applet

- Select **One mean** from the Scenario pull-down menu
- You can check Paste data to copy and paste in the raw data, or type in the sample size, mean, and standard deviation. Press **Calculate**.
- Check the box for **Test of significance** and enter the hypothesized value of  $\mu$  and set the direction of the alternative. Press **Calculate**.
- Check the box for **Confidence interval**, enter the confidence level and press **Calculate CI**.

Scenario: One mean

n:	130
mean, $\bar{x}$ :	98.249
sample sd, s:	.733
<input checked="" type="checkbox"/> Test of significance	
$H_0: \mu =$	98.6
$H_a: \mu \neq$	98.6

- (y) Verify your by-hand calculations with technology and summarize the conclusions you would draw from this study (both from the p-value and the confidence interval, including the population you are willing to generalize to). Also include interpretations, in context, of your p-value and your confidence level.

## Study Conclusions

The sample data provide very strong evidence that the mean body temperature of healthy adults is not 98.6 degrees ( $t = -5.46$ , two-sided p-value  $< 0.001$ ). This indicates there is less than a 0.1% chance of obtaining a sample mean as far from 98.6 as 98.249 in a random sample of 130 healthy adults from a population with mean body temperature of 98.6°F. A 95% confidence interval for the population mean body temperature is (98.122, 98.376), so we can be 95% confident that the population mean body temperature among healthy adults is between 98.122 and 98.376 degrees. This interval is entirely less than 98.6, consistent with our having found very strong evidence to reject 98.6 as a plausible value for the population mean. We are 95% confident, meaning if we were to repeat this procedure on thousands of random samples, in the long-run roughly 95% of the resulting intervals would successfully capture the population mean. We believe these procedures are valid because the sample size of 130 should be large enough unless there is severe skewness or outliers in the population.

**Practice Problem 2.5A**

Explore the last statements in the above results box using  $t$  confidence intervals:

- (a) Continue with the [Simulating Confidence Intervals](#) applet. Explore the coverage rate of the  $t$ -procedure with random samples from an Exponential (skewed) population for  $n = 5$ ,  $n = 100$ , and  $n = 200$ . Assess and summarize the performance of this  $t$ -procedure in each case.
- (b) Repeat (a) for a Uniform population distribution with endpoints  $a = 80$  and  $b = 85$ .

**Practice Problem 2.5B**

A study conducted by Stanford researchers asked children in two elementary schools in San Jose, CA to keep track of how much television they watch per week. The sample consisted of 198 children. The mean time spent watching television per week in the sample was 15.41 hours with a standard deviation of 14.16 hours.

- (a) Carry out a one-sample  $t$ -test to determine whether there is convincing evidence that average amount of television watching per week among San Jose elementary children exceeds fourteen hours per week. (Report the hypotheses, test statistic, p-value, and conclusion at the 0.10 level of significance.)
- (b) Calculate and interpret a one-sample 90%  $t$ -confidence interval for the population mean.
- (c) Comment on whether the technical conditions for the  $t$ -procedures are satisfied here. [Hint: What can you say based on the summary statistics provided about the likely shape of the population?]

### Investigation 2.6: Healthy Body Temperatures (cont.)

Reconsider the research question of Investigation 2.5 where a “one-sample  $t$ -test” found convincing evidence with sample mean body temperature of  $\bar{x} = 98.249^{\circ}\text{F}$  and a sample standard deviation of  $s = 0.733^{\circ}\text{F}$  against the hypothesis that  $\mu = 98.6$  in the population. In fact, we were 95% confident that  $\mu$  actually fell between 98.122 and 98.376 degrees (so not all that far from 98.6).

- (a) So if you recorded a body temperature of  $98.6^{\circ}$  would you be convinced you were sick? What temperature values would you be concerned about?
  
  
  
  
  
- (b) Is it true that approximately 95% of the temperatures in the sample fall inside your confidence interval? What percentage does? Do you think 95% of the temperatures in the population fall within the confidence interval?

### Prediction Intervals

It is very important to keep in mind that the confidence interval reported above only makes statements about the population mean, not individual body temperatures. But what if we instead wanted an interval to estimate the body temperature of a *single* randomly selected healthy adult rather than the population *mean* body temperature?

- (c) Still using our same sample results, what one number would you use to predict a healthy body temperature? If you then considered the uncertainty (margin-of-error) in this estimate for predicting one person’s body temperature, would you expect this margin-of-error to be larger or smaller than for predicting the population mean body temperature? Explain.

Estimate:

Margin-of-error:

Explain:

To construct such a confidence interval (often called a “prediction interval” to indicate that it will predict an individual outcome rather than the population mean), we need to consider both the sample-to-sample variation in sample means as well as the individual-to-individual variation in body temperatures.

- (d) We will estimate this combined standard error of an individual value by  $s\sqrt{1+1/n}$ . Using this formula, how will this compare to the standard error of the sample mean (larger or smaller)? Explain.

- (e) Calculate this value for the body temperature data.

If we are willing to assume that the population follows a normal distribution (note this is more restrictive than what we need to apply the Central Limit Theorem), then we can construct a 95% prediction interval for an individual outcome using the *t*-distribution.

**Definition:** To predict an individual value, we can calculate a [prediction interval](#) (PI). We construct the interval using the sample mean as our estimate, but we adjust the standard error to take into account the additional variability of an individual value from the population mean:

$$\bar{x} \pm t^*_{n-1} \sqrt{s^2 + s^2/n} \quad \text{or} \quad \bar{x} \pm t^*_{n-1} \times s \sqrt{1 + 1/n}$$

This procedure is valid as long as the sample observations are randomly selected from a normally distributed population. Note that prediction intervals are *not* robust to violations from the normality condition even with large sample sizes.

(f) Notice the critical value will be the same as in the previous investigation. Recall or determine the critical value with  $n = 130$  and 95% confidence.

(g) Using your answer to (f), calculate a 95% prediction interval for an individual healthy adult body temperature.

(h) How do the center and width of this interval compare to the 95% confidence interval for the population mean body temperature found in the previous investigation?

(i) Provide a one-sentence interpretation on the interval calculated in (g).

(j) The *JAMA* article only reported the summary statistics and did not provide the individual temperature values. If you had access to the individual data values, what could you do to assess whether the normality assumption is reasonable?

(k) Without access to the individual data values but considering the context (body temperatures of healthy adults), do you have any thoughts about how plausible it is that the population is normally distributed?

**Study Conclusions**

For one person to determine whether they have an unusual body temperature, we need a prediction interval rather than a confidence interval for the population mean. A 95% prediction interval for the body temperature of a healthy adult turns out to be (96.79, 99.71), a fairly wide interval. We are 95% confident that a randomly selected healthy adult will have a body temperature in this interval. This prediction interval procedure is valid only if population of healthy body temperatures follows a normal distribution, which seems like a reasonable assumption in this context

**Practice Problem 2.6A**

- (a) Examine the expression for a confidence interval for a population mean  $\mu$ . What happens to the half-width of the interval as the sample size  $n$  increases? Describe its limiting behavior.
- (b) Examine the expression for a prediction interval for an individual observation. What happens to the half-width of the interval as the sample size  $n$  increases? Describe its limiting behavior.
- (c) Explain why the differences in your answers to (a) and (b) make sense.

**Practice Problem 2.6B**

Recall the Stanford study on television viewing habits from Practice Problem 2.5B, with a sample size of 198, sample mean 15.41 and sample standard deviation 14.16.

- (a) Use this information to calculate a 95% prediction interval.
- (b) Provide a one-sentence interpretation of this interval.
- (c) Do you think this interval procedure is valid for these data? Explain.

## Summary of One-sample $t$ Procedures

**Parameter:**  $\mu$  = the population mean

**To test  $H_0: \mu = \mu_0$**

Test statistic:  $t_0 = (\bar{x} - \mu_0) / (s / \sqrt{n})$

Degrees of freedom =  $n - 1$

**$t$ -Confidence interval for  $\mu$ :**  $\bar{x} \pm t^*_{n-1} \times s / \sqrt{n}$

**Technical conditions:** These procedures are considered valid if the sample distribution is reasonably symmetric or the sample size is at least 30.

### Technology

- **R:** `t.test(data, hypoth, alt="greater", "less", or "two.sided", conf.level)` or ← raw data  
`iscamonesamplet(xbar, sd, n, hypothesized, alternative, conf.level)` ← summary data

- **Theory-Based Inference applet**

Use the pull-down menu to select **One Mean**. Specify the summary statistics (the sample size, sample mean, and sample standard deviation  $s$ ) or paste in the raw data. Check the box for test of significance and specify the hypothesized value, use the button to specify the direction of the alternative ( $<$  for not equal) and press Calculate and/or check the box for confidence interval, specify the confidence level and press Calculate CI.

## Summary Prediction Intervals for Individual Observations

**Prediction Interval:**  $\bar{x} \pm (t^*_{n-1}) \times s \sqrt{1 + 1/n}$

Valid only with normal population of observations

## SECTION 3: INFERENCE FOR OTHER STATISTICS

The previous section illustrated some very nice mathematical results for the distribution of the sample mean when sampling from an infinite process or a finite population. However, we also saw in Section 1 that the mean may not be the parameter we are most interested in and/or may not even be possible to calculate depending on the form of the data. In this section, you will look at a possible inference procedure when the median is of interest, as well as a newer methodology that works for any statistic of your choosing.

### Investigation 2.7: Water Oxygen Levels

Scientists often monitor the “health” of water systems to determine the impact of different changes in the environment. For example, Riggs (2002) reports on a case study that monitored the dissolved oxygen downstream from a commercial hog operation. There had been problems at this site for several years (e.g., manure lagoon overflow), including several fish deaths in the previous three years just downstream of a large swale through which runoff from the hog facility had escaped. The state pollution control agency decided to closely monitor dissolved oxygen downstream of the swale for the next three years to determine whether the problem had been resolved. In particular, they wanted to see whether there was a tendency for the dissolved oxygen level in the river to be less than the 5.0 mg/l standard. Sampling was scheduled to commence in January of 2000 and run through December of 2002. The monitors took measurements at a single point in the river, approximately six tenths of a mile from the swale, once every 11 days.

(a) Identify the observational units in this study. Would you consider this sampling from a population or from a process?

(b) Because the dissolved oxygen measurements were taken in the same location at fixed time intervals, would you consider this a simple random sample? Do you think the sample is likely to be representative of the river conditions? Explain.

**Definition:** A systematic sample takes selects observations at fixed intervals (e.g., every 10<sup>th</sup> person in line). If the initial observation is chosen at random and there is no structure in the data matching up to the interval size (e.g., every 7<sup>th</sup> day), then such samples are generally assumed to be representative of the population. In fact, we will often simplify the analysis by assuming they behave like simple random samples.

(c) Examine the data from the first year in [WaterQuality.txt](#). Describe the shape, center, and variability of the distribution. In particular, how do the mean and median compare? Do these data appear to be well-modelled by a normal distribution?

(d) State the null and alternative hypotheses for testing whether the long-run mean dissolved oxygen in this river is less than 5.0 mg/l (indicating too little oxygen in the water, causing problems in the aquatic community structure).

(e) Is the one-sample *t*-test likely to be valid for these data? Explain why or why not.

(f) An alternative analysis involves recoding the observations as “compliant” and “not compliant.” If we say a measurement is non-compliant when the dissolved oxygen is below 5.0 mg/l, how many non-compliant measurements do we have in this data set? What proportion of the sample is below 5.0 mg/l? [Hint: What do you want to do with the observation that is equal to 5.0?]

(g) Carry out a test for deciding whether the proportion of measurements that fall below 5.0 mg/l is statistically significantly larger than 0.50.

(h) If you decide that more than half of the time the river is non-compliant (that is, more than 50% of measurements are below 5.0 mg/l), what does that imply about the long-run *median* dissolved oxygen level? Explain.

(i) Identify one advantage to this analysis over the one-sample *t*-test. Identify one disadvantage of this analysis over the one-sample *t*-test.

- (j) The researchers actually wanted to decide whether the river was non-compliant more than 10% of the time. How would that change your analysis in (g) and would the p-value be larger or smaller?

### Study Conclusions

Both the mean (not shown) and of the median indicate that dissolved oxygen in this river tended to below the 5.0 mg/l that was cited as the “action level.” Although we don’t have statistically significant evidence that the long-run median DO level is below 5.0 (binomial p-value  $\approx 0.30$ ), the researchers actually wanted to engage in remedial action if the river is found to be in non-compliance significantly more than 10% of the time. The exact binomial probability of observing 19 or more non-compliant values from a process with a 0.10 chance of non-compliance is  $4.16 \times 10^{-11}$ , leaving “little doubt that the 10% non-compliant criterion was exceeded at the monitoring site during 2000.”

**Discussion:** When we use 0.50 as the hypothesized probability of success and we count how many of our quantitative observations exceed some pre-specified level, this is called the [sign test](#) and corresponds to a test of whether the population median equals that pre-specified level. The sign test can be advantageous over the *t*-test if the validity conditions for the *t*-test are not met (e.g., skewness in sample, not a large sample size). However, although this procedure focuses on how often you are above or below that level, it does not provide information about *how far* above or below as a confidence interval for the mean could.

### Practice Problem 2.7

Return to Practice Problem 2.2A. Carry out a test to determine whether there is convincing evidence that the median student estimate of 30 seconds differs from 30. (State your hypothesis, statistic and/or test statistic, and p-value.)

### Investigation 2.8: Turbidity

Another measure of water quality is turbidity, the degree of opaqueness produced in water by suspended particulate matter. Turbidity can be measured by seeing how light is scattered and/or absorbed by organic and inorganic material. Larger nephelometric turbidity units (NTU) indicate increased turbidity and decreased light penetration. If there is too much turbidity, then not enough light may be penetrating the water, affecting photosynthesis to the surface and leading to less dissolved oxygen. Riggs (2002) provides 244 turbidity monthly readings ([turbidity.txt](#)) that were recorded between 1980-2000 from a reach of the Mermentau River in Southwest Louisiana. The unit of analysis was the monthly mean turbidity (NTU) computed from each month's systematic sample of 21 turbidity measurements. The investigators wanted to determine whether the mean turbidity was greater than the local criterion value of 150 NTU.

- (a) Verify that a log-normal probability model is reasonable for these data.
- (b) Carry out a log transformation and report the mean, median, and standard deviation of the transformed data.
- (c) Are the mean and median values in (b) similar to each other? Is the mean of the logged turbidity values similar to the log of the mean of the turbidity values? Is the median of the logged turbidity values similar to the log of the median of the turbidity values?

Original scale:

$$\text{mean} =$$

$$\text{median} =$$

$$\log(\text{mean}) =$$

$$\log(\text{median}) =$$

Transformed scale:

$$\text{mean} =$$

$$\text{median} =$$

- (d) Explain why the  $\text{median}(\log(\text{turbidity}))$  is expected to be the same as  $\log(\text{median}(\text{turbidity}))$ , but this interchangeability is not expected to work for the mean.

- (e) Use a one-sample  $t$ -confidence interval to estimate the mean of the log-turbidity value for this river.

- (f) Back-transform the endpoints of this interval to return to the original units.

**Discussion:** Once you create a symmetric distribution, the mean and median will be similar. However, although transforming the data does not affect the ordering of the observations, it does impact the scaling of the values. So whereas the median of the transformed data is equal to taking the log of the median of the original data (at least with an odd number of observations), this does not hold for the mean. So when we back-transform our interval for the center of the population distribution, we will interpret this in terms of the median value rather than the mean value.

### Study Conclusions

A 95% confidence interval based on the transformed data equals (4.20, 4.40). Therefore, we will say we are 95% confident that the *median* turbidity in this river is between 66.69 NTU and 81.45 NTU, clearly less than the 150 NTU regulation. However, another condition for the validity of this procedure is that the observations are independent. Further examination of these data reveals seasonal trends. Adjustments need to be made to account for the seasonality before these data are analyzed.

### Practice Problem 2.8

Return to the [honking.txt](#) data from Investigation 2.2.

- (a) Use the log transformation to calculate a 95% confidence interval for the mean log response time.
- (b) Back-transform and interpret the interval.
- (c) Outline a method for verifying that this procedure results in a 95% confidence interval for the population median.

### Investigation 2.9: Heroin Treatment Times

Hesketh and Everitt (2000) report on a study by Caplehorn and Bell (1991) that investigated the times that heroin addicts remained in a clinic for methadone maintenance treatment. The data in [heroin.txt](#) includes the amount of time that the subjects stayed in the facility until treatment was terminated (*times*, number of days). For about 37% of subjects, the study ended while they were still in the clinic (status = 0). Thus, their “survival time” has been “truncated.” For this reason, we might not want to focus on the mean survival time, but rather some other measure of a “typical” survival time. We will explore both the median and the 75<sup>th</sup> percentile. We will treat this group of 238 patients as representative of the population of heroin addicts.

- (a) Produce and describe a histogram of the survival times for these patients.

*R hint:* We will just focus on the *times* data, so you can copy and load all the data and then let  
`> times = heroin$times`

- (b) Explain why is not likely that we could find a simple mathematical model for this distribution.

Even though a sample size of  $n = 238$  is likely enough for us to employ the Central Limit Theorem for the distribution of the sample mean, we may be more interested in the median or another statistic like a “trimmed mean.” However, we may not know a mathematical model for the sampling distribution of one of these other statistics. Of course, what we are most interested in is a measure of the sample-to-sample variability of that statistic. Previously, we estimated the standard error of the sample mean by using information from the sample ( $s/\sqrt{n}$ ). But when such a formula does not exist, another method that is gaining in popularity is *bootstrapping*.

**Definition:** A [bootstrap sample](#) resamples the data from the existing sample, drawing the same number of observations, but *with replacement*. A bootstrap distribution is a collection of values of the statistic from lots of bootstrap samples.

The reasoning behind this technique is that if the sample has been randomly selected, it should be representative of the population. Thus, it should give us information about the population and about samples drawn from that population. In other words, rather than assume a particular probability model for the population, bootstrapping assumes that the population looks just like the sample, replicated infinitely many times. By sampling with replacement from the original sample, and calculating the statistics of interest for each bootstrap sample, we gain information about the shape and spread of the sampling distribution of the statistic of interest.

**Key result:** The *plug-in principle* states that the standard deviation of the bootstrap distribution provides a reasonable estimate of the standard deviation of the sampling distribution of the statistic.

(c) Use technology to take a bootstrap sample from the heroin data.

- In R:

```
> sample1 = sample(238, times, replace = TRUE)
```

Now calculate the median of your bootstrap sample. (Notice we could calculate any statistic we wanted.)

(d) Take another bootstrap sample. Did you get the same sample median? Why or why not?

To estimate the sample to sample variation in our statistic (e.g., the median), we now want to repeat this process a large number of times.

(e) Use technology to create a bootstrap distribution from the heroin data.

### Technology Detour – Bootstrap Sampling

In R create a list of functions and then apply those functions using the median as the statistic

```
> bootstrapmedians = boot(x, samplemedian, R = 1000)
```

OR install the *boot* package and write a function to calculate the median of a sample and use

```
> resamples = lapply(1:1000, function(i) sample(times,
      replace=T))
> bootstrapmedians = sapply(resamples, median)
```

You should now have a column of all the medians from 1000 different bootstrap samples.

(f) Produce graphical and numerical summaries of the 1000 bootstrap medians. Describe the behavior of the bootstrap distribution.

(g) Suggest a quick method for using the distribution in (d) to calculate a rough confidence interval for the population median. *Hint:* Recall our usual  $\text{estimate} \pm 2 \times \text{standard deviation of estimate}$ . Calculate and interpret this interval.

**Discussion:** Once you get an estimate of the standard deviation of the statistic, there is still some consideration of how to construct a confidence interval. In this case, the distribution of medians should look fairly symmetric, so a symmetric confidence interval can work. In other cases, there are more complicated “adjustments” that can be made to get a bootstrap confidence interval. Discussion of these methods is beyond of the scope of this text, but you have seen an example of the power of bootstrapping

– you can create a distribution *for any statistic* that should have the same variability as the sampling distribution of that statistic, without having to pretend you know anything about the population!

(h) Repeat this analysis, but now using the upper quartile (75% percentile) as the statistic.

- **In R:**

```
> bootstrapquantiles = sapply(resamples, quantile)
> upper75 = bootstrapquantiles[4L:4L, ]
```

### Study Conclusions

The distribution of treatment times was skewed to the right with a median of about 1 year (367.5 days) and interquartile range of 418.5 days. Because of the skewness and truncation in the data, we might prefer the median or even the third quartile as the statistic. Based on a bootstrap simulation, an approximate 95% confidence interval for the median treatment time in the population of heroin addicts is between 306 days and 430 days (but results will vary), while an approximate 95% confidence interval for the upper quartile is around 534 to 638 days (results will vary). The interval for the median includes lower values than a *t*-interval for the population mean (368, 437), because the *t*-procedure is more strongly affected by the skewness in the data. We would like to generalize these data to all heroin patients, but would like more information to insure that this sample is representative.

### Practice Problem 2.9A

- Apply this method to obtain an estimate of the standard deviation of the 25% trimmed mean. [Hints: Create and sort a sample, then subset the middle 50% of the values and find the mean of those values.]
- What is an advantage to the trimmed mean compared to the mean?
- What is an advantage to the trimmed mean over the median?

### Practice Problem 2.9B

Return to the [honking.txt](#) data from Investigation 2.2.

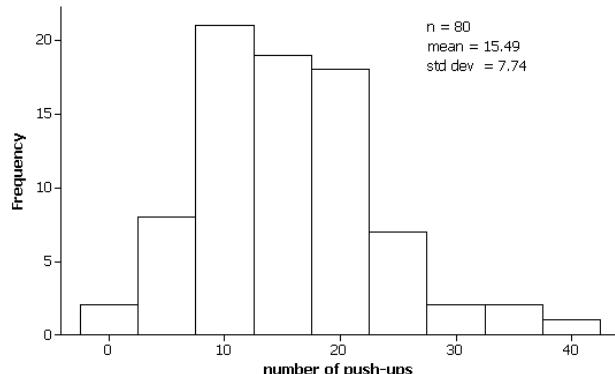
- Suggest a reason why the Central Limit Theorem may not apply with these data.
- Create and interpret an informal ( $\pm 2SE$ ) bootstrap confidence interval for the population median.

**Example 2.1: Pushing On**

*Try these questions yourself before you use the solutions following to check your answers.*

As in many states, California mandates physical fitness testing at different grade levels. The recommended number of push-ups for 12 year old males is 10-20 and for 13 year old males is 12-25. A sample of 80 7<sup>th</sup> grade males was obtained at a rural high school in Central California (Wetzel and Hernandez, 2004). Data was gathered using the measurement techniques defined by the state. (The feet are together, hands will be shoulder width apart, the subjects back will be straight, and their eyes will be looking toward the horizon. The arms need to bend to a 90-degree angle, while keeping their back flat. The push-ups are counted at a set tempo without stopping to rest.)

A histogram and summary statistics of the sample data are below:



(a) Describe the distribution of the results in this sample.

(b) Suppose we take random samples of 80 males from a very large population. According to the Central Limit Theorem, what can you say about the behavior of the sampling distribution of the sample means calculated from these samples?

(c) If this were a random sample from a population, would the sample data provide strong evidence that the population mean differs from 20 push-ups? Conduct a significance test to address this question. Also calculate confidence intervals to estimate the population mean with various levels of confidence.

(d) What precautions should we have in analyzing these results?

- (e) Would it be reasonable to use this sample to calculate a prediction interval for the number of push-ups by a 7<sup>th</sup> grade male in California?

## Analysis

- (a) For this sample, the distribution of the number of push-ups completed is slightly skewed to the right. The average number of push-ups by the 80 males in the sample was 15.49, with standard deviation 7.74 push-ups. Most students completed about 10-25 push-ups, with the maximum around 40 or so.
- (b) Because the sample size is large ( $80 > 30$ ), the sampling distribution of sample means should be approximately normal, regardless of the population shape, with mean  $\mu$ , and standard deviation equal to  $\sigma/\sqrt{80}$ . The exact values of  $\mu$  (the population mean number of push-ups done by 7<sup>th</sup> grade males) and  $\sigma$  (the population standard deviation) are unknown, but they should be in the ball park of 15.49 and 7.74, the sample statistics.
- (c) *Test of significance:* If this was a random sample from a larger population of 7<sup>th</sup> graders, let  $\mu$  represent the mean number of push-ups that would be completed in this population. We want to decide whether  $\mu$  is significantly different from 20.

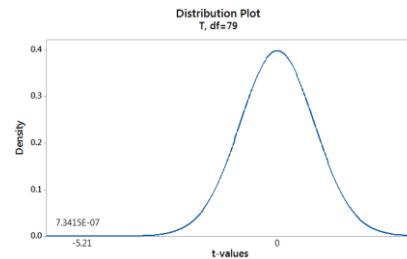
$$H_0: \mu = 20 \text{ (the population mean number of push-ups is 20)}$$

$$H_a: \mu \neq 20 \text{ (the population mean differs from 20)}$$

Because we are working with a quantitative response variable and the sample size is large, we will model the sampling distribution of the standardized statistic with the *t* distribution with  $80 - 1 = 79$  degrees of freedom.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{15.49 - 20}{7.74/\sqrt{80}} = -5.21$$

$$\text{p-value} = 2P(T_{79} \leq -5.21) = 2(0.0000007) = 0.0000014$$

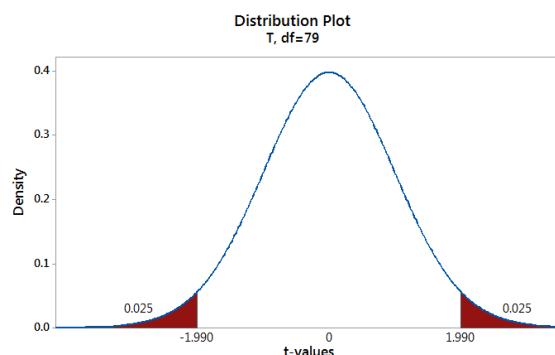


With such a small p-value, we easily reject the null hypothesis and conclude that the population mean number of push-ups differs from 20.

**Confidence Intervals:**

To construct a 95% confidence interval for  $\mu$ , we will use  $t^*_{79} = 1.990$

$$\bar{x} \pm t^*_{n-1} s / \sqrt{n} = 15.49 \pm 1.990(7.74 / \sqrt{80}) = 15.49 \pm 1.72 = (13.76, 17.21)$$



Verifying these calculations in Minitab, we would find:

Test of mu = 20 vs not = 20

N	Mean	StDev	SE Mean	95% CI	T	P
80	15.4900	7.7400	0.8654	(13.7675, 17.2125)	-5.21	0.000

Based on this sample, assuming it represents the larger population, we are 95% confident that the average number of push-ups completed by all 7<sup>th</sup> graders in the population is between 13.76 and 17.21, so 20 is rejected as a plausible value at the 0.05 level of significance. We could also find 99% and 99.9% confidence intervals for  $\mu$  to be:

$$95\%: \bar{x} \pm t^*_{n-1} s / \sqrt{n} = 15.49 \pm 2.640(7.74 / \sqrt{80}) = 15.49 \pm 2.28 = (13.21, 17.77)$$

$$99\%: \bar{x} \pm t^*_{n-1} s / \sqrt{n} = 15.49 \pm 3.418(7.74 / \sqrt{80}) = 15.49 \pm 2.96 = (12.53, 18.45)$$

Thus, even with these stricter standards of 99% and 99.9% confidence, we still have reason to believe that the population mean is less than 20 push-ups. These results are consistent with the extremely small p-value from the significance test above.

(d) We should be very cautious in generalizing these results as we don't know if the push-up performance of the students sampled at this rural high school in Central California is representative of a larger population 7<sup>th</sup> graders in the state. The margin of error calculated only takes into account the expected amount of random sampling error, not any biases from our sampling methods.

(e) If we wanted to calculate a prediction interval, we would have the same concern that this sample may not be representative of 7<sup>th</sup> graders across the state, and the additional concern that the population distribution may not follow a normal distribution. We have reason to doubt that it does since the sample shows some skewness to the right. So using these data to predict the number of push-ups by an individual in this population would be risky.

**Example 2.2: Distracted Driving?**

*Try these questions yourself before you use the solutions following to check your answers.*

Strayer and Johnston (2001) asked student volunteers to use a machine that simulated driving situations. At irregular intervals, a target would flash red or green. Participants were instructed to press a “brake button” as soon as possible when they detected a red light. The machine would calculate the reaction time to the red flashing targets for each student in milliseconds.

The students were given a warm-up period to familiarize themselves with the driving simulator. Then the researchers had each student use the driving simulation machine while talking on a cell phone about politics to someone in another room and then again with music or a book-on-tape playing in the background (control). The students were randomly assigned as to whether they used the cell phone or the control setting for the first trial. The reaction times (in milliseconds) for 16 students appear below and in the file [driving.txt](#).

Subject	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Cell	636	623	615	672	601	600	542	554	543	520	609	559	595	565	573	554
Control	604	556	540	522	459	544	513	470	556	531	599	537	619	536	554	467

(a) Suppose we let  $X$  represent the number of subjects whose cell phone reaction time was longer. What probability distribution can we use to model  $X$ ? What assumptions are behind this probability model?

(b) How could you characterize the null and alternative hypotheses in this case (in symbols and in words)?

(c) What is the observed value of  $X$  in this study? How would you determine whether this statistic provides convincing evidence against the null hypothesis?

(d) Use the binomial distribution to determine a p-value for a sign test applied to this study. Interpret this p-value in the context of this study and state the conclusions you would draw from this p-value.

## Analysis

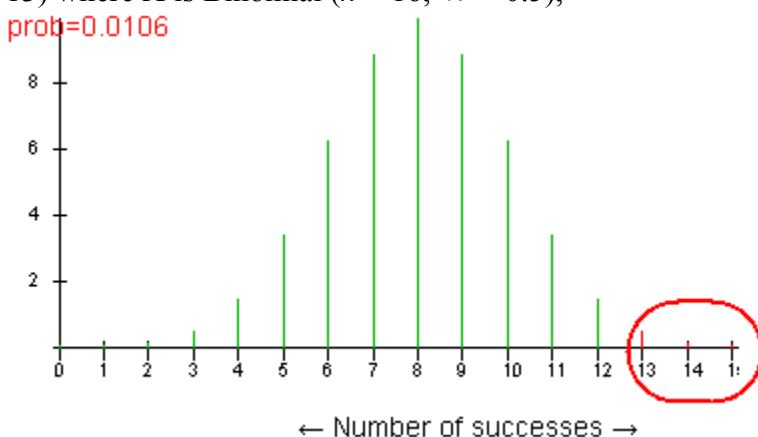
(a) We can use the binomial distribution to model  $X$ , where  $n = 16$  (the number of drivers in the study) and  $\pi$  represents the probability that the cell phone reaction time is longer. This is a valid model as long as the reaction times between subjects are independent and the probability of the cell phone reaction time being longer is the same across the subjects.

(b)  $H_0: \pi = 0.5$  (equally likely for either reaction time to be longer)

$H_a: \pi > 0.5$  (more likely for a subject to take longer to react when talking on the cell phone).

(c) The observed value of  $X$  is 13 in this study. So we need to assess how unusual it is for 13 or more drivers out of 16 to have a longer cell phone reaction time under the null hypothesis that there is no long-run tendency for either reaction time to be longer.

(d) Determining  $P(X \geq 13)$  where  $X$  is Binomial ( $n = 16$ ,  $\pi = 0.5$ ),



With an exact Binomial p-value of 0.0106, we would reject the null hypothesis at the 5% level of significance. Only 1.1% of random assignments (of which reaction time goes to which condition for each person) would have at least 13 of the differences being positive if the null hypothesis were true. We have strong evidence that such a majority did not happen by chance alone, but reflects a genuine tendency for cell phone reaction times to be longer.

## CHAPTER 2 SUMMARY

In this chapter, you focused on distributions of quantitative data, following up earlier discussion in Investigation A. In exploring quantitative data, we start with a graph displaying the distribution, such as a dotplot or a histogram. The most relevant features include shape, center, and variability, as well as deviations from the overall pattern. The appropriate numerical summaries of center include the mean and median, but keep in mind that the median is more resistant than the mean to outliers and skewness. The appropriate numerical summaries of variability include interquartile range (IQR) and standard deviation, with the IQR also more resistant to outliers and skewness. Remember when examining the *variability* of a data set, you are interested in the spread of the values that the variable takes (along the horizontal axis). In summarizing data, the five-number summary can be very descriptive and you were also introduced to boxplots, a graphical display of the five-number summary.

After examining the sample distribution, it may be appropriate to ask whether a sample statistic differs significantly from some claim about the population. For example, whether the sample mean far enough from a conjectured population mean to convince you that the discrepancy did not arise “by chance alone.” You explored some simulation models, mostly to convince you that the *t*-distribution provides a very convenient approximation to the sampling distribution of the standardized statistic of a sample mean. This allowed us to approximate p-values and calculate confidence intervals. Keep in mind that the logic of statistical significance hasn’t changed – we assume the null hypothesis is true and determine how often we would get a sample statistic at least as extreme as what was observed. You saw that this likelihood is strongly affected by not only the sample size but also how much variability is inherent in the data. You also learned about prediction intervals, as opposed to confidence intervals, for estimating individual values rather than the population mean.

These *t* procedures do come with some validity conditions and should be applied with extreme caution with small sample sizes or extreme skewness in the distribution. In this case, you could consider transforming the variable to rescale the observations to a known probability distribution such as the normal distribution. You can also apply the *sign test* which looks at the sign of the values rather than the numerical outcome and allows you to apply the binomial distribution. Other approaches such as bootstrapping are very flexible in the choice of statistic.

## SUMMARY OF WHAT YOU HAVE LEARNED IN THIS CHAPTER

- Numerical and graphical summaries for describing the distribution of a quantitative variable
  - Focus on shape, center, and variability of the distribution, and deviations from overall pattern
- Properties of numerical summaries (resistance)
- How to describe, measure, and interpret variability of a quantitative variable
- Identifying and considering potential explanations for outliers
- Stating hypotheses in terms of a population mean
- One-sample *t*-procedures
  - When appropriate to apply; Why helpful
  - How to interpret results
  - Alternative approaches if not valid (e.g., data transformations)
- Factors that affect the significance of the comparison (e.g., sample size, sample variability)
- Prediction intervals

**TECHNOLOGY SUMMARY**

- Summarizing data (dotplots, histograms, boxplots, mean, standard deviation, five-number summary)
- Simulating a sampling distribution from a finite population
- One-sample *t*-procedures
- Transforming data

**Choice of Procedures for Analyzing One Mean**

<b>Study design</b>	One quantitative variable from large population or process	
<b>Descriptive</b>	<u>R Commands</u> <code>iscamdotplot</code> <code>histogram</code> (lattice package) <code>iscamboxplot</code> or <code>boxplot</code> <code>iscamsummary</code> <code>qqplot</code>	<u>Minitab Commands</u> <code>Graph &gt; Dotplot</code> <code>Graph &gt; Histogram</code> <code>Graph &gt; Boxplot</code> <code>Stat &gt; Basic Statistics &gt; Display Descriptive Statistics</code> <code>Graph &gt; Probability Plot</code>
<b>Parameter</b>	$\mu$ = population mean or process mean (or use sign test or bootstrapping)	
<b>Null Hypothesis</b>	$H_0: \mu = \mu_0$	
<b>Simulation</b>	Random sample from finite population or bootstrapping	
<b>Can use <i>t</i> procedures if</b>	Normal population (symmetric sample distribution) or large sample size (e.g., larger than 30)	
<b>Test statistic</b>	$t_0 = (\bar{x} - \mu) / (s / \sqrt{n})$	
<b>Confidence interval</b>	$\bar{x} \pm t_{n-1}^* \times s / \sqrt{n}$	
<b>Prediction interval</b>	$\bar{x} \pm t_{n-1}^* \times s \sqrt{1 + 1/n}$ (with normally distributed population)	
<b>R Commands</b>	<code>t.test(x, alternative="two.sided", mu=0, conf.level = 0.95)</code> raw data in "x" <code>iscamonesamplet</code> <a href="#">summary statistics</a> <ul style="list-style-type: none"> <li>• <code>xbar</code> (<math>\bar{x}</math>, sample mean)</li> <li>• <code>sd</code> (<math>s</math>, sample standard deviation)</li> <li>• <code>n</code> (sample size)</li> <li>• <code>hypothesized mean</code> (<math>\mu_0</math>)</li> <li>• <code>alternative</code> ("less", "greater", or "two.sided")</li> <li>• <code>Optional: conf.level(s)</code></li> </ul>	
<b>Minitab</b>	<code>Stat &gt; Basic Statistics &gt; 1-Sample t</code> Input variable column or summary statistics	

## CHAPTER 3: COMPARING TWO PROPORTIONS

In this chapter, you will focus on comparing results from two groups on a categorical variable. These groups could be samples from different populations or they could have been deliberately formed during the design of the study (a *third* source of possible randomness). You will again consider multiple ways to analyze the statistical significance of the difference in the groups, namely simulation, exact methods, and normal approximations to answer whether the observed difference in the groups could have happened “by chance alone.” You will also continue to consider issues of statistical confidence and types of errors. A key consideration to keep in mind will be the scope of conclusions that you can draw from the study based on how the data were collected.

### Section 1: Comparing two population proportions

- Investigation 3.1: Teen hearing loss (cont.) – Tables, conditional props, bar graphs,  $z$ -procedures
- Investigation 3.2: Nightlights and near-sightedness – Association, confounding

### Section 2: Types of Studies

- Investigation 3.3: Handwriting and SAT scores – Observational studies, experiments
- Investigation 3.4: Have a nice trip – Random assignment, scope of conclusions
- Investigation 3.5: Botox for back pain – Designing experiments

### Section 3: Comparing two treatment probabilities

- Investigation 3.6: Dolphin therapy – Randomization test
- Investigation 3.7: Is yawning contagious? – Fisher’s exact test
- Investigation 3.8: CPR vs. chest compressions –  $z$ -procedures

### Section 4: Other Statistics

- Investigation 3.9: Flu vaccine – Relative risk
- Investigation 3.10: Smoking and lung cancer – Types of observational studies, odds ratio
- Investigation 3.11: Sleepy drivers – Application

Example 3.1: Wording of Questions

Example 3.2: Worries about Terrorist Attacks

## SECTION 1: COMPARING TWO POPULATION PROPORTIONS

### Investigation 3.1: Teen Hearing Loss (cont.)

The Shargorodsky, Curhan, Curhan, and Eavey (2010) study from Investigation 1.16 actually focused on comparing the current hearing loss rate among teens (12-19 years) to previous levels to see whether teen hearing loss is increasing, possibly due to heavier use of ear buds. In addition to the 1771 participants in the NHANES 2005-6 study (333 with some level of hearing loss), they also had hearing loss data on 2928 teens from NHANES III (1988-1994), with 480 showing some level of hearing loss. Our goal is to assess whether the difference between these two groups can be considered statistically significant.

(a) What is the primary difference between this study and those examined in Chapter 1? What is the same?

(b) Identify the two populations and the variable being considered in this study.

Populations:

Variable:

Type:

### Descriptive Statistics

(c) When we have two samples with a categorical variable, it is often useful to organize the data using a two-way table. Complete the following  $2 \times 2$  table of counts.

	<b>1988-1994</b>	<b>2005-2006</b>	<b>Total</b>
<b>Some hearing loss</b>	480	333	
<b>No hearing loss</b>			
<b>Total</b>	2928	1771	

(d) Explain why it does not make sense to conclude that hearing loss was more prevalent in 1988-1994 than in 2005-2006 based only on the comparison that  $480 > 333$ .

(e) Suggest a better way to compare the prevalence of hearing loss between the two studies. Calculate *one number* as the statistic for this study (what symbols are you using?). Does your statistic seem large enough to convince you that there has been an increase in hearing loss?

**Definition:** The simplest statistic for comparing a binary variable between two groups is the difference in the proportion of “successes” for each group. These proportions, calculated separately for each group rather than looking at the overall proportion, are called conditional proportions.

In this case, we compute the difference in the *proportion* of teens with some level of hearing loss between the two years ( $\hat{p}_{94} = 480/2928 - \hat{p}_{06} = 333/1771$ ).

The next step is to examine an effective graphical summary for comparing the two groups.

### Technology Detour – Segmented Bar Graphs

#### In R

- Create (and view) a matrix in R to store the counts from the two-way table
 

```
> hearing = matrix(c(480, 2448, 333, 1438), nrow=2,
+ dimnames = list(c("some loss", "no loss"), c("94", "06"))))
```

```
> hearing
```
- Convert to a matrix of conditional proportions (margin =1 for row proportions, margin = 2 for column proportions)
 

```
> hearingprop = prop.table(hearing, margin=2)
```

```
> hearingprop
```
- Create a segmented bar graph
 

```
> barplot(hearingprop, legend=T, ylab="proportion")
```

(f) Use technology to create numerical and graphical summaries for these summarized data. Write a sentence or two comparing the distributions of hearing loss between these two studies. Be sure to report an appropriate statistic.

(g) Is it possible that the two sample proportions (with some hearing loss) might have differed by this much even if the overall population proportion was the same for both years?

## Inferential Statistics

The previous question should look very familiar. As we've said before, it certainly is *possible* to obtain *sample* proportions this far apart, just by random chance, even if the *population* proportions (of teens with some hearing loss) were the same. The question now is to ask *how likely* such a difference would be if the population proportions were the same. We can answer this question by modeling the sampling variability, arising from taking random samples from these populations, for the difference in two sample proportions. Investigating this sampling variability will help us to assess whether this particular difference in sample proportions is strong evidence that the population proportions actually differ.

- (h) Let  $\pi_{94}$  represent the proportion of all American teenagers in 1994 with at least some hearing loss, and similarly for  $\pi_{06}$ . Define the parameter of interest to be  $\pi_{94} - \pi_{06}$ , the difference in the population proportions between these two years. State appropriate null and alternative hypotheses about this parameter to reflect the researchers' conjecture that hearing loss by teens is becoming more prevalent.

$$H_0:$$

$$H_a:$$

- (i) Explain in your own words what we want the p-value to measure in this case.

## Simulation

- (j) As before, we will begin our inferential analysis by assuming the null hypothesis is true. What "random process" are we simulating? What is the source of the randomness in this study? What do we need to assume to perform this simulation?

- (k) So under the null hypothesis we really only have one value of  $\pi$  to estimate – the common population proportion with hearing loss for these two years. What is your best estimate for  $\pi$  from the sample data? [Hint: Think about combining the two years together.]

- (l) Describe how you could carry out a simulation analysis to investigate whether the observed difference in sample proportions provides strong evidence that the population proportions with hearing loss differed between these two time periods. [Hint: Think about replicating the randomness in the study but in a world where the null hypothesis is true.]

We will begin our simulation analysis by assuming the population proportion is actually this value ( $\pi = (480+333)/(2928+177) = 0.173$ ). We simulate the drawing of two different random samples from this population, one to represent the 1994 study and the other for the 2006 study. Because the population is very large compared to the sample sizes, we will model this by treating the population as infinite and sampling from a binomial process. Then we examine the distribution of the difference in the conditional proportions with some hearing loss between these two years. Finally, we repeat this random sampling process for many trials. [Note: We can assume  $\pi = 0.173$  without loss of generality, but you might want to verify this with other values for  $\pi$  as well.]

(m) Follow the commands below to randomly sample one “could have been” difference in conditional proportions *under the null hypothesis*:

1. Simulate taking one random sample of 2928 teenagers (representing the 1994 study) from a population in which  $\pi = 0.173$  have at least some hearing loss and count the number of successes (with hearing loss in the sample):

*n = 2928, common  $\pi = 0.173$*

- **In R**

```
> count1994 = rbinom(1, size = 2928, prob=.173)
> count1994
```

*Displays the results*

2. Now simulate (and view) taking a random sample of 1771 teenagers (for the 2006 study) from a population also with  $\pi = 0.173$  and count the number of successes in this sample.

*Technology:* Repeat the above steps but change the sample size and column/vector name.

3. Calculate (by hand) the difference in the conditional proportions with hearing loss for these two “could have been” samples. [*Hint:* How do you do this based on the randomly generated “number of successes” in each sample?]

(n) Will everyone in the class get the same answers to (m)? Explain.

(o) How does the difference in sample proportions you just generated at random, assuming that the two years have the same population proportion, compare to the observed difference in sample proportions in the actual studies from (e)?

SIMILAR

FURTHER FROM ZERO

CLOSER TO ZERO

But to evaluate how unusual such a result is when the null hypothesis is true, we want to generate many more outcomes assuming the null hypothesis to be true.

(p) To create the null distribution of the differences in sample proportions, you will generate 1000 random samples from each population and calculate the difference in conditional proportions each time, storing the results in 2 different vectors/columns:

**Number of random samples = 1000**• **In R**

```
> phat94 = rbinom(1000, 2928, .173)/2928
> phat06 = rbinom(1000, 1771, .173)/1771
> phatdiffs = phat94 - phat06
```

Divides number of successes by  $n$   
to get sample proportions ( $X/n$ ).  
You now have 1000 differences.

(q) Display a histogram and summary statistics of the null distribution of differences in sample proportions:

• **In R**

```
> par(mfrow=c(3,1))
> hist(phat94); hist(phat06)
> hist(phatdiffs, labels=T)
> mean(phat94); mean(phat06)
> mean(phatdiffs)
> sd(phat94); sd(phat06)
> sd(phatdiffs)
```

Creates 3 rows in graph window  
Use a semicolon to separate commands  
In RStudio, can zoom the graph window

Describe the null distribution of the difference in sample proportions. Does it behave as you expected? In particular, does the mean of this distribution make sense? (How does it compare to the means of the individual  $\hat{p}$  distributions?) Explain.

(r) Now determine the empirical p-value by counting how often the simulated difference in conditional proportions is at least as extreme as the actual value observed in the study:

These commands count how many simulated differences are below the observed statistic ( $-0.024$ ) by creating a Boolean variable (1 when the statement is true, 0 otherwise), summing the ones, and then dividing by the number of samples in the simulation.

• **In R**

```
> pvalue=sum(phatdiffs <= -.024)/1000
> pvalue
```

Report your empirical p-value and indicate what conclusion you would draw from it.

## Mathematical Model

(s) It turns out that there is no “exact” method for calculating the p-value here, because the difference in two binomial variables does not have a binomial, or any other known, probability distribution. However, did the histogram you examined remind you of any other probability distribution?

(t) Overlay a normal curve on your null distribution and/or examine a normal probability plot to evaluate whether the simulated differences appear to “line up” with observations from a standard normal distribution.

- In R
 

```
> iscamaddnorm(phatdiffs)
> qqnorm(phatdiffs)
```

Does the normal model appear to be a reasonable approximation to the null distribution?

## Probability Detour

There is a theoretical result that the *difference* in two normal distributions will also follow a normal distribution. When our sample sizes ( $n_1$  and  $n_2$ ) are large, we know the individual binomial distributions are well approximated by normal distributions. Consequently, the difference of the sample proportions will be well approximated by a normal distribution as well. The mean of this distribution is simply the difference in the means of the individual normal distributions.

(u) How does the variability (SD) of the *difference* in  $\hat{p}$  values compare to the variability of the individual  $\hat{p}$  distributions? Explain why this makes intuitive sense.

SIMILAR

LARGER

SMALLER

Explanation:

### Central Limit Theorem for the difference in two sample proportions

When taking two independent samples (of sizes  $n_1$  and  $n_2$ ) from large populations, the distribution of the difference in the sample proportions ( $\hat{p}_1 - \hat{p}_2$ ) is approximately normal with mean equal to  $\pi_1 - \pi_2$  and standard deviation equal to  $SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$ .

Under the null hypothesis  $H_0: \pi_1 - \pi_2 = 0$ , the standard deviation simplifies to  $\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

where  $\pi$  is the common population proportion.

**Technical Conditions:** We will consider the normal model appropriate if the sample sizes are large, namely  $n_1\pi_1 \geq 5$ ,  $n_1(1-\pi_1) \geq 5$ ,  $n_2\pi_2 \geq 5$ ,  $n_2(1-\pi_2) \geq 5$ , and the populations are large compared to the sample sizes.

**Note:** The variability in the differences in sample proportions is larger than the variability of individual sample proportions. In fact, the *variances* (standard deviation squared) add, and then we take the square root of the sum of variances to find the standard deviation.

However, to calculate these values we would need to know  $\pi_1$ ,  $\pi_2$ , or  $\pi$ . So again we estimate the standard deviation of our statistic using the sample data.

**Case 1:** When the null hypothesis is true, we are assuming the samples come from the same population, so we “pool” the two samples together to estimate the common population proportion of successes.

That is, we estimate  $\pi$  by looking at the ratio of the total number of successes to the total sample size:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\text{total number of successes}}{\text{total sample size}}$$

Then we use this value to calculate the standard error of  $\hat{p}_1 - \hat{p}_2$  to be:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

(v) Use these theoretical results to suggest the general *formula* for a test statistic and a method for calculating a p-value to test  $H_0: \pi_1 - \pi_2 = 0$  (also expressed as  $H_0: \pi_1 = \pi_2$ ) versus the alternative  $H_a: \pi_1 - \pi_2 < 0$ . (This is referred to as the [two-sample z-test](#) or [two proportion z-test](#).)

$$\text{Test statistic } z = \frac{\text{observed-hypothesized}}{\text{standard error}} =$$

(w) Calculate and interpret the value of the test statistic specified in (v) as applied to the hearing loss study.

$$\hat{p}_{94} - \hat{p}_{06} = \hat{p} =$$

$$SE(\hat{p}_{94} - \hat{p}_{06}) =$$

$$\text{test statistic: } z =$$

interpretation:

(x) Is the standard error close to the empirical standard deviation from your simulation results?

(y) Compute the p-value for this test statistic and compare it to your simulation results.

## Confidence Interval

**Case 2:** When we make no assumptions about the populations (for example, when we are not testing a particular null hypothesis but only estimating the parameter), we will use a different formula to approximate the standard deviation of  $\hat{p}_1 - \hat{p}_2$ .

(z) Using the above CLT result, suggest a general formula for a confidence interval for the difference in population proportions  $\pi_1 - \pi_2$ .

(aa) Calculate and interpret a 95% confidence interval to compare hearing loss of American teenagers in these two years. Is this confidence interval consistent with your test of significance? Explain.

**Note:** It is technically incorrect to say there has been a 0.1% to 4.7% increase, because “percentage change” implies a multiplication of values, not an addition or subtraction as we are considering here. It would be acceptable to say that the increase is between 0.1 and 4.7 *percentage points*.

## Technical Conditions

The above Central Limit Theorem holds when the populations are much larger than the samples (e.g., more than 20 times the sample size) and when the sample size is large. We will consider the latter condition met when we have at least 5 successes and at least 5 failures in each sample (so there are four numbers to check).

**Note:** A “Wilson adjustment” can be used with this confidence interval as before, this time putting one additional success and one additional failure in each sample. This adjustment will be most useful when the proportions are close to 0 or 1 (that is when the sample size conditions above are not met).

(bb) Summarize your conclusions from this study. Be sure to address statistical significance, statistical confidence, and the populations you are willing to generalize the results to. Also, are you willing to conclude that the change in the prevalence of hearing loss is due to the increased use of ear buds among teenagers between 1994 and 2006? Explain why or why not.

## Study Conclusions

We have moderate evidence against the null hypothesis ( $p\text{-value} \approx 0.02$ , meaning we would get a difference in sample proportions  $\hat{p}_1 - \hat{p}_2$  as small as  $-0.024$  or smaller in about 2% of random samples from two populations with  $\pi_1 = \pi_2$ ). We are 95% confident that the population proportion with some hearing loss is between 0.001 and 0.047 higher “now” than ten years ago. We feel comfortable drawing these conclusions about the populations the NHANES samples were selected from as they were random samples from each population (and there was no overlap in the populations between these two time periods). However, there are many things that have changed during this time period, and it would not be reasonable to attribute this increase in hearing loss exclusively to the use of ear buds.

## Practice Problem 3.1A

- When we conducted the simulation analysis above, we used the same probability of “success” (having some hearing loss) for both years. Why did we do this?
- How did we decide what common probability of success to use?
- Why did we count how many samples of the simulation gave a result of  $-0.024$  or smaller (explain the  $-0.024$  part and the “or smaller” part).

## Technology Detour – Two-sample z-procedures

### Theory-Based Inference applet

- Select **Two proportions**
- Check the box to paste in 2 columns of data (stacked or unstacked) and press **Use data** or specify the sample sizes and either the sample counts or the sample proportions and press **Calculate**.
- For the test, check the box for **Test of Significance** Keep the hypothesized difference at zero and set the direction of the alternative, press **Calculate**.
- For the confidence interval, check the box, specify the confidence level and press **Calculate CI**

### In R

- Use `iscamtwopropztest` which takes the following inputs:
  - observed1* (either the number of successes or sample proportion for first group), *n1* (sample size for first group), *observed2* (count or proportion), and *n2*
  - Optional: hypothesized difference* and *alternative* (“less”, “greater”, or “two.sided”)
  - Optional: conf.level*

For example: > `iscamtwopropztest(480, 2928, 333, 1771, 0, alt = "less", conf.level= 95)`

finds the p-value for a one-sided alternative as well as a 95% confidence interval for  $\pi_1 - \pi_2$ .

## Summary of simulation and z-procedures for comparing two population proportions

**Parameter:**  $\pi_1 - \pi_2$  = the difference in the population proportions of success

**To test  $H_0: \pi_1 - \pi_2 = 0$ :**

1. *Simulation:* Random samples from binomial processes with common  $\pi$

2. *Two-sample z-test:*

The test statistic  $z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$  where  $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{\text{total number of successes}}{\text{total number in study}}$

is well approximated by the standard normal distribution when  $n_1 \hat{p} \geq 5$ ,  $n_1(1 - \hat{p}) \geq 5$ ,  $n_2 \hat{p} \geq 5$  and  $n_2(1 - \hat{p}) \geq 5$ , where  $\hat{p}$  is the overall proportion of successes in the two groups put together.

**Approximate  $(100 \times C)\%$  Confidence interval for  $\pi_1 - \pi_2$ :**

1. *Two-sample z-interval:*

An approximate  $(100 \times C)\%$  interval:  $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$

where  $-z^*$  is the  $100 \times (1 - C)/2$ <sup>th</sup> percentile from the standard normal distribution.

This method is considered valid when  $n_1 \hat{p}_1 \geq 5$ ,  $n_1(1 - \hat{p}_1) \geq 5$ ,  $n_2 \hat{p}_2 \geq 5$ , and  $n_2(1 - \hat{p}_2) \geq 5$ .

2. Or the *Wilson adjustment* can be made on the sample proportions and sample sizes first.

**Note:** We can simplify the technical conditions by verifying that you have independent random samples from two processes or large populations (or one random sample classified by a binary variable) and at least 5 successes and 5 failures in each group.

**Technology:** To perform the calculations for these two-sample z-procedures you can use the Theory-Based Inference applet (two proportions) or Minitab (Stat > Basic Statistics > 2 Proportions) or iscamtwopropztest in R.

### Practice Problem 3.1B

*USA Today* reported that newspapers appear to be losing credibility over time (March, 2004). They cited a nationwide sample of 1,002 adults, 18 years or older, interviewed via telephone (under the direction of Princeton Survey Research Associates) during the period May 6-16, 2002. Of the 932 respondents who were able to rate the daily newspaper they were most familiar with, 587 rated the paper as “largely believable” (a 3 or a 4 on the scale). When the same question was asked four years earlier (May 7-13, 1998), 922 said they could rate their daily paper and of those, 618 rated the paper as “largely believable.”

- (a) Are the technical conditions met to calculate the two-sample z-interval for these data? Explain.
- (b) Calculate and interpret a 95% confidence interval from these data.

### Investigation 3.2: Nightlights and Near-sightedness

Myopia, or near-sightedness, typically develops during the childhood years. Recent studies have explored whether there is an association between development of myopia and the use of night-lights with infants. Quinn, Shin, Maguire, and Stone (1999) examined the type of light children aged 2-16 were exposed to. Between January and June 1998, the parents of 479 children who were seen as outpatients in a university pediatric ophthalmology clinic completed a questionnaire (children who had already developed serious eye conditions were excluded). One of the questions asked was “Under which lighting condition did/does your child sleep at night?” before the age of 2 years. The following two-way table classifies the children’s eye condition and whether or not they slept with some kind of light (e.g., a night light or full room light) or in darkness.

	Some light	Darkness	Total
Near-sighted	188	18	206
Not near-sighted	119	154	273
Total	307	172	479

**Definition:** When we have two variables, we often specify one to be the [explanatory variable](#) and the other to be the [response variable](#). The explanatory variable is the one that we think might be influencing or explaining changes in the response variable, the outcomes of interest.

- (a) Which variable, lighting condition or eye condition, would you consider the explanatory variable in this study and which the response variable?
- (b) Calculate conditional proportions to measure the difference in the rate of near sightedness between the children with and without lighted rooms in this sample. (Use appropriate symbols.) Also produce a segmented bar graph to compare the eye condition of the two lighting groups. Comment on what this graph reveals about differences in eye condition between these lighting groups. Do you think this difference will be statistically significant? Explain.

**Discussion:** In the previous investigation, the researchers literally took two different random samples from two different groups of teens. In this study, the data arose from one sample and each child in the sample was classified according to two variables: *lighting condition* and *eye condition*. A natural research question here might be “Does use of nightlights and room lights increase the rate of near-sightedness in children?” In such a situation, we can instead phrase our hypotheses in terms of the *association* between the two variables:

$H_0$ : There is no association between *lighting condition* and *eye condition*

$H_a$ : There is an association between *lighting condition* and *eye condition*

We can also state these hypotheses in terms of underlying population probabilities:

$H_0: \pi_{\text{light}} - \pi_{\text{darkness}} = 0$       vs.  $H_a: \pi_{\text{light}} - \pi_{\text{darkness}} > 0$   
 where  $\pi_i$  represents the probability of a near-sighted child in population  $i$ .

- (c) Briefly discuss how you could modify the simulation procedure used in Investigation 3.1 to reflect the changes in the study design.

It turns out, it will be valid to use the same inferential analysis procedures here as in Investigation 3.1 as long as the samples (e.g., room light, darkness) can be considered independent of each other. One counter example would be if the observational units had been *paired* in some way (e.g., brothers and sisters). But if we don't believe the responses of particular children in lit rooms in this study in any way relate to the responses of particular children in the dark rooms, we consider the samples *independent*, even though they weren't literally sampled separately, and use the same Central Limit Theorem to specify a normal distribution as a model of the sampling distribution of the difference in the two sample proportions.

- (d) Use technology to perform a two-sample  $z$ -test to compare the proportion with near sightedness between these two groups. Report the test statistic and p-value. Also report and interpret a 95% confidence interval from these data.

- (e) Because the p-value is so small, would you be willing to conclude that the use of lights *causes* an increase in the probability of near-sightedness in children? Explain. If not, suggest a possible alternative explanation for the significantly higher likelihood of near sightedness for children with lighting in their rooms.

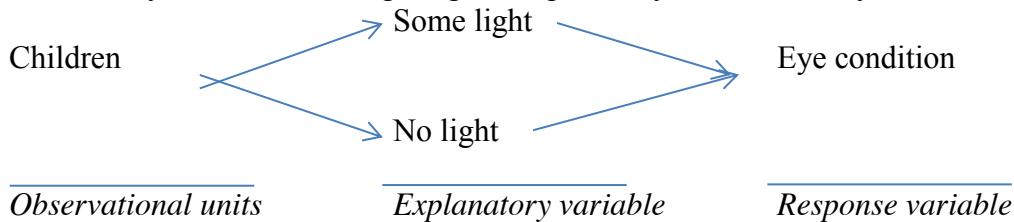
**Discussion:** When we find a highly significant difference between two groups, although we can conclude a significant *association* between the explanatory and response variables, we are not always willing to draw a *cause-and-effect* conclusion between the two variables. Though many wanted to use this study as evidence that it was the lighting that caused the higher rate of near-sightedness, for all we know it could have been children with poorer vision that asked for more light to be used in their rooms.

Or there could even be a third variable that is related to the first two and could provide the real explanation for the observed association. Such a variable is called a *confounding variable*.

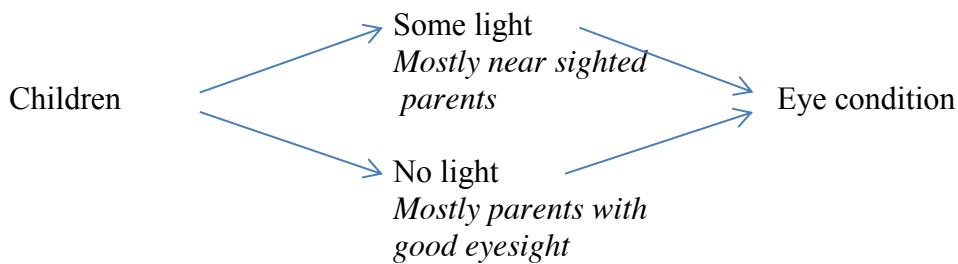
**Definition:** In some studies, we may consider one variable the *explanatory variable* that we suspect has an effect on the *response variable*. However, in some studies often there may be other confounding variables that also differ between the explanatory variable groups *and* have a potential effect on the response variable.

Your alternative explanation in (d) probably identified a potential confounding variable. For example, *eye condition of the parents* may be related to the explanatory variable (parents with poorer sight are more likely to use room lighting) and to the response variable (children of near-sighted parents are more likely to be near-sighted themselves). Because this confounding variable changes with the explanatory variable, we have no way of separating out the effects of this variable and those of the explanatory variable. This prevents us from drawing a cause-and-effect conclusion between *lighting condition* and *eye condition*.

One way to visualize confounding is to examine a diagram of the study design. For the night-light and eye condition study we considered lighting the explanatory variable and eye condition the response.



But we are conjecturing the parents' eye condition might also be changing across the two groups.



So when those who had lighting in their room develop near-sightedness, there are two equally plausible explanations because two things have changed for that group and we can't separate them out.

(f) It has been suggested that “low birth weight” is a possible confounding variable in this study. Does this seem plausible to you? Describe circumstances under which this would be a confounding variable.

(g) In this study, 70% of the children were Caucasian. Does this mean “race” is a possible confounding variable? What should you be concerned about? What about the region of the country where the clinic was located?

**Discussion:** Keep in mind that the issue of confounding, and whether or not you can draw a cause-and-effect conclusion between explanatory and response variables, is very different from the issue of generalizability. We are willing to generalize to a larger population when the study design involves *random sampling* of observational units from the population. This is a separate question from how the observational units were formed into subgroups.

### Study Conclusions

With a very small p-value, we have extremely strong evidence that the relationship we observed in the sample data (use of light in the child’s room is associated with higher rates of myopia) did not occur by chance alone. There appears to be a genuine connection here, more than luck of the draw, but we can’t necessarily draw a causal relationship between lighting and eyesight from this study. We should also be cautious in generalizing these results too far beyond the type of families that visit this clinic in this part of the United States.

**Practice Problem 3.2A**

For each of the following statements,

- (i) identify which is being considered the explanatory variable and which the response variable,
- (ii) then suggest a potential confounding variable that explains the observed association between the explanatory and response variables,
- (iii) finally explain how the suggested variable is related both to the response and to differences between the explanatory variable groups.

- (a) Children with larger feet tend to have higher reading scores than children with smaller feet.
- (b) Days with a higher number of ice cream sales also tend to have more drownings.
- (c) Cities with higher teacher salaries also tend to have higher sales volumes in alcohol.
- (d) People who eat apples regularly tend to have fewer cavities.
- (e) Some professional sports teams have better winning percentages when their home games are not sold out than when their home games are sold out.

**Practice Problem 3.2B**

A study (referred to in the October 12, 1999 issue of *USA Today*) reported the number of deaths by heart attack in each month over a 12-year period in New York and Boston. December and January were found to have substantially higher numbers of deaths than the other months. Researchers conjectured that the stress and overindulgence associated with the holiday season might explain the higher numbers of deaths in those months compared to the rest of the year.

- (a) Identify a confounding variable that provides an alternative explanation for this result. Make sure you indicate how this confounding variable affects the two explanatory variable groups differently.
- (b) Explain why saying “people living in big cities have more stressful lives” does not identify a confounding variable for analyzing the association between time of year and heart attacks.
- (c) A similar study conducted in Los Angeles also found more heart attack deaths in December and January than in other months. Identify a potential confounding variable from the New York and Boston studies that is controlled for in the Los Angeles study. Also identify a variable that is still potentially confounding in the Los Angeles study.
- (d) Suggest a way of altering the study (e.g., another location to examine) that would help isolate the effect of the holiday season.

## SECTION 2: TYPES OF STUDIES

In the previous investigation, we cautioned against jumping to a cause-and-effect conclusion when a statistically significant association is found between the response and explanatory variables. But when can we potentially draw such a conclusion?

### Investigation 3.3: Handwriting and SAT Scores

An article about handwriting appeared in the October 11, 2006 issue of the *Washington Post*. The article mentioned that among students who took the essay portion of the SAT exam in 2005-06, those who wrote in cursive style scored *significantly* higher on the essay, on average, than students who used printed block letters. Researchers wanted to know whether simply writing in cursive would be a way to increase scores.

- (a) Identify the explanatory and response variables in this study. Also classify the variables as quantitative or categorical.

Explanatory:

Response:

- (b) Would you conclude from this study that using cursive style *causes* students to score better on the essay? If so, explain why. If not, identify a potential confounding variable, and explain how it provides an alternative explanation for why the cursive writing group would have a significantly higher average essay score.

The article also mentioned a different study in which the same one essay was given to all graders. But some graders were shown a cursive version of the essay and the other graders were shown a version with printed block letters. Researchers randomly decided which version the grader would receive. The average score assigned to the essay with the cursive style was *significantly* higher than the average score assigned to the essay with the printed block letters.

- (c) Explain a key difference between how this study was conducted and how the first study was conducted.

- (d) Would the difference cited in (c) make you more willing to draw a cause-and-effect conclusion between writing style and SAT score with the second study than with the first one? Explain.

**Definition:** An [observational study](#) is one in which the researchers passively observe and record information about the observational units. In an [experimental study](#), the researchers actively impose the explanatory variable (often called *treatment*) on the observational units (can also be called the experimental units in an experiment).

(e) Classify the first and second study above as observational or experimental:

Study of 2006 exams:

Study of one response written both ways:

(f) Explain why the experimental study should be less susceptible to the potential for a confounding variables between the writing style and the score.

(g) Discuss one disadvantage to the experimental study compared to the observational study for this research question.

### Study Conclusions

The second study determined for each judge whether they would grade the exam written in block letters or in cursive letters. Because it was the same exam otherwise, there shouldn't be any other reason for the difference in scores apart from the type of writing. For this reason, because the difference was statistically significant, we will be willing to conclude that the writing style *caused* the difference in scores. However, this only tells us about that one exam and in the slightly artificial condition of asking the judges to grade the exam. We don't know whether the same effect would be found on other essays or conditions. By looking at the 2006 exams, we know that actual student papers were graded by actual judges, a more realistic setting, but also with the potential for confounding variables between the writing style and grade, such as student preparation prior to the exam.

### Practice Problem 3.3

In a study published in the July 2003 issue of the journal *Psychosomatic Medicine*, researchers reported that people who tend to think positive thoughts catch a cold less often than those who tend to think negative thoughts. The scientists recruited 334 initially healthy male and female volunteers aged 18 to 54 years through advertisements in the Pittsburgh area, and they first interviewed them over a two week period to gauge their emotional state, eventually assigning them a numerical score for positive emotions and a numerical score for negative emotions. Then the subjects were quarantined before the researchers injected rhinovirus, the germ that causes colds, into each subject's nose. The subjects were then monitored for five days for the development of cold-like symptoms. Subjects scoring in the bottom third for positive emotions were three times more likely to catch a cold than those scoring the top third for positive emotions. According to the above definitions, is this an observational study or a randomized experiment? Would it be valid to draw a cause and effect conclusion from their results? Explain.

### Investigation 3.4: Have a Nice Trip

An area of research in biomechanics and gerontology concerns falls and fall-related injuries, especially for elderly people. Recent studies have focused on how individuals respond to large postural disturbances (e.g., tripping, induced slips). One question is whether subjects can be instructed to improve their recovery from such perturbations. Suppose researchers want to compare two such recovery strategies, lowering (making the next step shorter, but in normal step time) and elevating (using a longer or normal step length with normal step time). Subjects will have first been trained on one of these two recovery strategies, and they will be asked to apply it after they feel themselves tripping. The researchers will then induce the subject to trip while walking (but harnessed for safety), using a concealed mechanical obstacle.

Suppose the following 24 subjects have agreed to participate in such a study. Both males and female were recruited because females tend to have better balance (lower center of gravity).

Females: Alisha, Alice, Betty, Martha, Audrey, Mary, Barbie, Anna

Males: Matt, Peter, Shawn, Brad, Michael, Kyle, Russ, Patrick, Bob, Kevin, Mitch, Marvin, Paul, Pedro, Roger, Sam

- (a) One way to design an experiment for this study would be to assign the eight females to use the elevating strategy and the 16 males to use the lowering strategy. Would this be a reasonable strategy? If not, identify a better method for deciding who uses which strategy.

**Definition:** In a well-designed experiment, experimental units are randomly assigned to the treatment groups. Each unit is equally likely to be assigned to any of the treatments.

- (b) Let's explore this random assignment process to determine whether it does "work." First, let's focus on the sex variable. Suppose we put each person's name on a slip, put those slips in a hat and mix them up thoroughly, and then randomly draw out 12 slips for names of people to assign to the elevating strategy. What proportion of this group do you expect will be male? What proportion of the lowering strategy do you expect will be male? Do you think we will always get a 8/8 split of the male subjects (8 males in each treatment group)?

(c) To repeat this random assignment a large number of times to observe the long-run behavior, we will use the [Randomizing Subjects](#) applet. Open the applet, and press the **Randomize** button.

### Randomizing 24 Subjects

Group 1		
name	sex	height

What proportion of subjects assigned to Group 1 are men? Of Group 2? What is the difference in these two proportions?

You will notice that the difference in the proportion male is shown in the dotplot in the bottom graph. In this graph, the observational unit is one repetition of the random assignment, and the variable is the difference in proportions of men between the two groups.

(d) Press the **Randomize** button again. Was the difference in proportions of men the same this time?

(e) Change the number of replications from 1 to 198 (for 200 total, and press the **Randomize** button. The dotplot will display the difference between the two proportions of men for each of the 200 repetitions of the random assignment process. Where are these values roughly centered?

(f) Click on the most extreme dot (positive or negative). The graph should update to show the groups. What difference in proportion male did you find here? Did you ever get an absolute difference as large as 0.6667 (a 12/4 split)? Which is more likely, ending up with a 12/4 split in males or ended up with an 8/8 split in males?

(f) Does random assignment *always* equally distribute/balance the men and women between the two groups?

(g) Is there a *tendency* for there to be a *similar* proportion of men in the two groups? How are you deciding? What does this tell you about the plausibility of any later difference in the two groups being attributed to females having better balance?

**Discussion:** A 12/4 split in males would be problematic, because then if the treatment group that had all males showed worse balance, we wouldn't know whether it was because of the strategy they use or because they were all males and the other group was more likely to recover their balance because they had mostly females. But with random assignment, such an unequal split is unlikely. Instead, the random assignment usually creates an 8/8 split or a 9/7 split. In this case, the groups are balanced on the sex variable and sex is no longer confounded with recovery.

(h) Prior research has also shown that the likelihood of falling is related to variables such as walking speed, stride rate, and height, so we would like the random assignment to distribute these variables equally between the groups as well. In the applet, use the pull-down menu to switch from the sex variable to the height variable. The dotplot now displays the differences in *average* height between Group 1 and Group 2 for these 200 repetitions. In the long-run, does random assignment tend to equally distribute the height variable between the two groups? Explain.

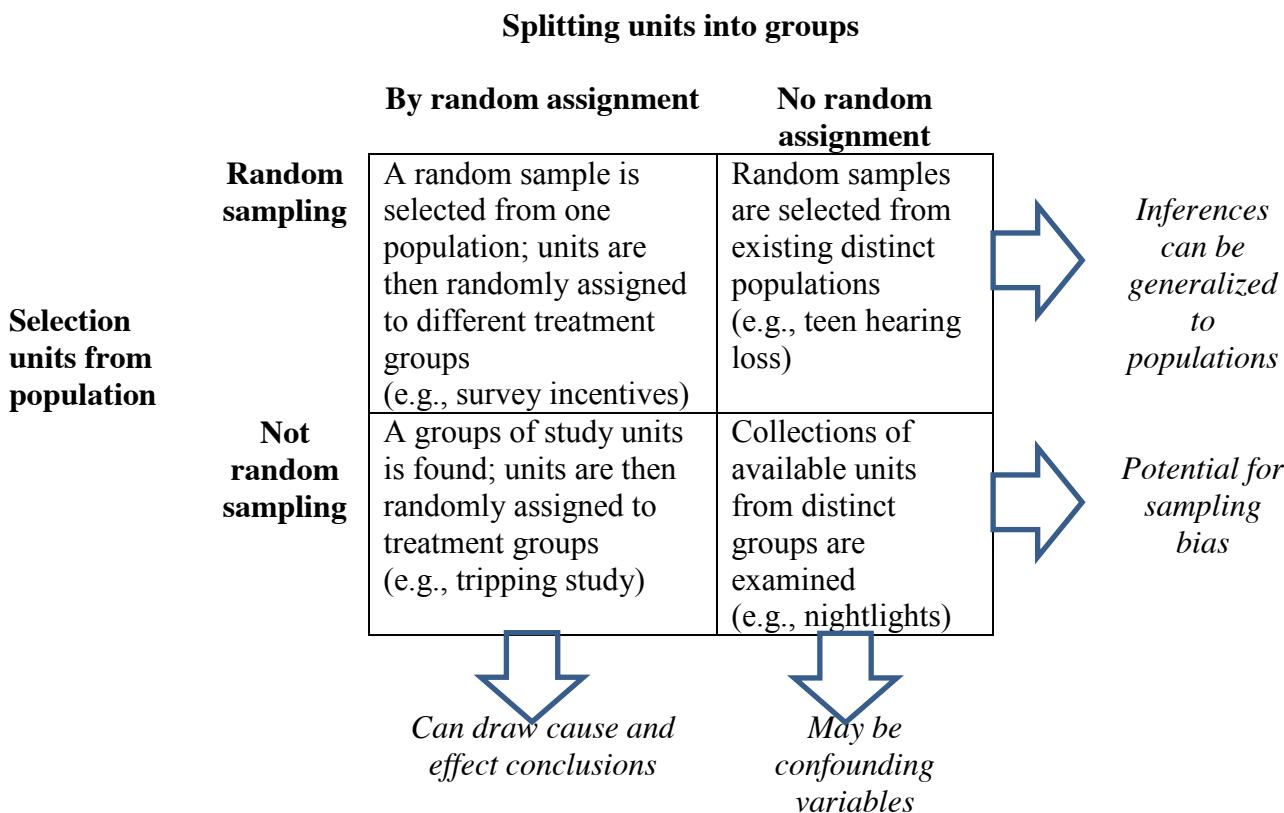
(i) Suppose there is a “balance gene” that is related to people’s ability to recover from a trip. We didn’t know about this gene ahead of time, but if you select the “Reveal gene?” button and then select “gene” from the pull-down menu, the applet shows you this gene information for each subject and also how the proportions with the gene differ in the two groups. Does this variable *tend* to equalize between the two groups in the long run? Explain.

(j) Suppose there were other “*x*-variables” that we could not measure such as stride rate or walking speed. Select the “Reveal both?” button and use pull-down menu to display the results for the *x*-variable (*x-var*). Does random assignment generally succeed in equalizing this variable between the two groups or is there a tendency for one group to always have higher results for the *x*-variable? Explain.

**Discussion:** The primary goal of *random assignment* is to create groups that equalize any potential confounding variables between the groups, creating explanatory variable groups that overall differ only on the explanatory variable imposed. Note that this “balancing out” applies equally well to variables that can be observed (such as sex and height) and variables that may not have been recorded (such as age or walking speed) or that cannot be observed (such as a hidden gene). Although we could have forced variables like sex and height to be equally distributed between the two groups, the virtue of random assignment is that it also tends to balance out variables that we might not have thought of ahead of time and variables that we might not be able to see or control. Thus, when we observe a “statistically significant” difference in the response variable between the two groups at the end of the study, we feel more comfortable attributing this difference to the explanatory variable (e.g., recovery strategy employed), because **that should have been the only difference between the groups**.

(k) Suppose the “tripping” study finds a statistically significant difference between the two strategies. What conclusion would you draw? For what population? What additional information would you need to know?

Keep in mind that cause-and-effect conclusions are a separate consideration from the generalizability of the results to a larger population. From now on we will consider each of these issues in our final conclusion as shown in the table below (adapted from Ramsey and Schafer's *The Statistical Sleuth*):



### Practice Problem 3.4

A team of researchers (Singer, et al., 2000) used the Survey of Consumer Attitudes to investigate whether incentives would improve the response rates on telephone surveys. A national sample of 735 households was randomly selected, and all 735 of the households were sent an “advance letter” explaining that the household would be contacted shortly for a telephone survey. However, 368 households were randomly assigned to receive a monetary incentive along with the advance letter, and the other 367 households were assigned to receive only the advance letter.

- Identify the explanatory and response variables.
- Is this an observational study or an experiment? Explain how you are deciding.
- If these researchers find a statistically significant difference in the response rate between these two groups, would it be reasonable to draw a cause-and-effect conclusion? Explain (including any additional information you would need to know to decide).
- To what population is it reasonable to generalize these results? Explain (including any additional information you would need to know to decide).

### Investigation 3.5: Botox for Back Pain

A 2001 study (Foster, Clapp, Erickson, & Jabbari, *Neurology*, “Botulinum toxin A and chronic low back pain: A randomized, double-blind study”) examined the efficacy of botulinum A (botox, used in various cosmetic and medical procedures) in lower back pain. An excerpt from the journal article follows:

Thirty-one consecutive patients with chronic low back pain who met the inclusion criteria were studied: 15 received 200 units of botulinum toxin type A, 40 units/site at five lumbar paravertebral levels on the side of maximum discomfort, and 16 received normal saline. Each patient’s baseline level of pain and degree of disability was documented using the visual analogue scale (VAS) and the Oswestry Low Back Pain Questionnaire (OLBPQ). The authors reevaluated the patients at 3 and 8 weeks (visual analogue scale) and at 8 weeks (OLBPQ).

At 3 weeks, 11 of 15 patients who received botulinum toxin (73.3%) had >50% pain relief vs four of 16 (25%) in the saline group ( $p = 0.012$ ). At 8 weeks, nine of 15 (60%) in the botulinum toxin group and two of 16 (12.5%) in the saline group had relief ( $p = 0.009$ ). Repeat OLBPQ at 8 weeks showed improvement in 10 of 15 (66.7%) in the botulinum toxin group vs three of 16 (18.8%) in the saline group ( $p = 0.011$ ). No patient experienced side effects.

(a) Was this an observational study or an experiment? Explain how you are deciding.

(b) Explain why it was important to use a comparison (“control”) group in this study (rather than simply giving botox to all subjects and then measuring whether they had experienced pain relief at the end of the study).

(c) Once a control group is formed, explain the purpose of using “normal saline.”

(d) Explain why it was important for the subjects not to know which treatment they were receiving.

(e) Did this study use random assignment or random sampling and why was this use of randomness important?

(f) Explain why standardized measurements of pain were used. Do you think the researchers should be the ones administering the measurement scales?

(g) Why is it important for the published article to outline the “inclusion criterion”?

**Definition:** A *randomized, comparative, double-blind experiment* includes two or more groups being compared, random assignment of “subjects” to groups, and double-blindness (neither the subjects nor the evaluators of the response variable knows which treatment group the subject is in, to guard against the [placebo effect](#) and other subjective biases). These types of studies, brought to the forefront of scientific attention by R.A. Fisher (1935), are considered the gold standard for determining cause and effect relationships between variables.

(h) For the night light and near-sightedness study, would it be feasible to conduct a randomized, comparative, double-blind experiment? Explain.

Keep in mind, it is not always feasible to design such a study. You also need to be cautious of the *Hawthorne effect* – people involved in a study sometimes act differently just by virtue of being in a study – and other issues of realism and feasibility. For example, how do we assess the long-term effects of second-hand smoking on pregnant women?

**Practice Problem 3.5A**

For each of the following research questions, describe how you would design an observational study to address the question and how you would design an experiment. In each case, identify which study you feel is more appropriate. In particular, are there any ethical or logistical issues that might prevent you from carrying out the experiment?

- (a) Are there effects of second-hand smoke on the health of children?
- (b) Do people tend to spend more money in stores located next to food outlets with pleasing smells?
- (c) Does cell phone use increase the rate of automobile accidents?
- (d) Do people consume different amounts of ice cream depending on the size of the bowl used?

**Practice Problem 3.5B**

Researchers conducted a randomized, double-blind trial to determine whether taking large amounts of Vitamin E protects against prostate cancer (*Journal of the National Cancer Institute*, 1998). To study this question, they enrolled 29,133 Finnish men, all smokers, between the ages of 50 and 69. The men were divided into two groups: One group took vitamin E and a second group took a placebo. The researchers followed all the men for eight years and then determined how many had developed prostate cancer. They found that participants taking vitamin E were significantly less likely to develop prostate cancer.

- (a) Explain what “randomized” means in this study and its purpose.
- (b) Explain what “double-blind” means in the context of this study and its purpose.
- (c) Explain what “significantly less likely” means in the context of this study.
- (d) Based on this report, is it reasonable to conclude that taking vitamin E *causes* a reduction in the probability of developing prostate cancer? Explain your reasoning.
- (e) Based on this report, what population is it reasonable to generalize these results to? Explain your reasoning.

### SECTION 3: COMPARING TWO TREATMENT PROBABILITIES

In the previous section, we saw potential benefits to using random assignment – creating groups that we are willing to consider equivalent. However, there is still the chance that “luck of the draw” could lead to a higher success proportion for the response variable in one group. In this section you will explore a method for calculating p-values that addresses how often random assignment could create a difference between the treatment groups at least as large as the one observed.

#### Investigation 3.6: Dolphin Therapy

Antonioli and Reveley (2005) investigated whether swimming with dolphins was therapeutic for patients suffering from clinical depression. The researchers recruited 30 subjects aged 18–65 with a clinical diagnosis of mild to moderate depression through announcements on the internet, radio, newspapers, and hospitals in the U.S. and Honduras. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in one hour of swimming and snorkeling each day, but one group (Dolphin Therapy) did so in the presence of bottlenose dolphins and the other group (Control) did not. At the end of two weeks, each subject’s level of depression was evaluated, as it had been at the beginning of the study, and each subject was categorized as experiencing substantial improvement in their depression symptoms or not. (Afterwards, the control group had one day session with dolphins.)

- (a) Identify the observational units and variables in this study. Also indicate which variable is being considered the explanatory variable and which the response variable.

Observational units:

Explanatory variable:

Response variable:

- (b) Was this an observational study or an experiment? Explain how you are deciding.

The following two-way table summarizes the results of this experiment:

	Dolphin Therapy	Control Group	Total
Showed substantial improvement	10	3	13
Did not show substantial improvement	5	12	17
Total	15	15	30

- (c) Calculate the difference in the conditional proportions of substantial improvement between the two explanatory variable groups.

- (d) Do the data appear to support the claim that dolphin therapy is more effective than the control program?

We must ask the same questions we have asked before – is it *possible* that this difference has arisen by random chance alone if there was no effect of the dolphin therapy? If so, how surprising would it be to observe such an extreme difference between the two groups?

One way to define a parameter in this case is to let  $\pi_{\text{dolphin}} - \pi_{\text{control}}$  denote the difference in the underlying probability of substantial improvement between the two treatment conditions.

- (e) State appropriate null and alternative hypotheses to reflect the researchers' conjecture in terms of this parameter.

$H_0$ :

$H_a$ :

- (f) So under the null hypothesis, we are assuming that there is no difference in the *treatment effect* between the two groups. In other words, whether or not people improved was not related to which group they were put in. Explain how you could design a simulation to help address this question, keeping in mind that this study involved *random assignment* not *random sampling*. Also keep in mind that this simulation will assume the null hypothesis is true.

One way to model this situation is by assuming 13 of the 30 people were going to demonstrate substantial improvement regardless of whether or not they swam with dolphins. Then the key question is **how unlikely is it for the random assignment process alone to randomly place 10 or more of these 13 improvers into the dolphin therapy group** (that is, a difference in the conditional proportions that improve of 0.467 or larger)? If the answer is that this observed difference would be very surprising if dolphin therapy was not more effective, then we would have strong evidence to conclude that dolphin therapy *is* more effective.

- (g) To model the chance variability inherent in the random assignment process, under the null hypothesis, take a set of 30 playing cards or index cards, one for each participant in the study. Designate 13 of them to represent improvements (e.g., red suited cards, blue colored index cards, or S-labeled index cards) and then mark 17 of them to represent non-improvements (e.g., black suited cards, green colored index cards, or F-labeled index cards). Shuffle the cards and deal out two groups: 15 of the “subjects” to the dolphin therapy group and 15 to the control group. Complete the following two-way table to show your “could have been” result under the null hypothesis.

Simulation Repetition #1:	Dolphin Therapy	Control Group	Total
<b>Showed substantial improvement</b>			13 (blue)
<b>Did not show substantial improvement</b>			17 (green)
<b>Total</b>	15	15	30

Also calculate the difference in the conditional proportions (dolphin – control) from this first simulated repetition, and indicate whether this simulated result is as extreme as the observed result:

(h) Repeat this hypothetical random assignment five times and record four more “could have been” tables:

Repetition 2:

	Dolphin	Control
Yes		
No		

Repetition 3:

	Dolphin	Control
Yes		
No		

Repetition 4:

	Dolphin	Control
Yes		
No		

Repetition 5:

	Dolphin	Control
Yes		
No		

**Note:** You may have noticed that once you entered the number of improvers in the Dolphin Group into the table, the other entries in the table were pre-determined due to the “fixed margins.” So for our statistic, we could either report the difference in the conditional proportions *or* the number of successes (improvers) in the Dolphin group as these give equivalent information.

(i) Pool your results for the *number of successes (blue index cards) assigned to the Dolphin Group* for your 5 repetitions with the rest of your class and produce a well-labeled dotplot of the null distribution.



(j) What are the observational units and variable in the above graph? [Hint: Think about what you would have to do to add another dot to the graph. How should you label the horizontal axis?]

(k) Granted, we have not done an extensive number of repetitions, but how's it looking so far? Does it seem like the actual experimental results (the observed 10/3 split) would be surprising to arise purely from the random assignment process under the null model that dolphin therapy is not effective? Explain.

We really need to do this simulated random assignment process hundreds, preferably thousands of times. This would be very tedious and time-consuming with cards, so let's turn to technology

(l) Open the [Dolphin Study](#) applet.

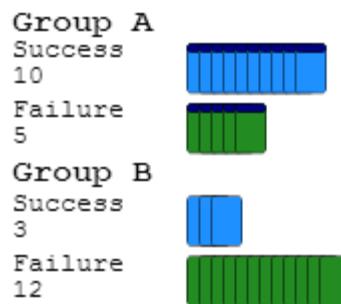
- Confirm that the two-way table displayed by the applet matches that of the research study.
- Check **Show Shuffle Options** and confirm that there are 30 cards, 13 blue and 17 green.
- Select the **Plot** option
- Press **Shuffle**.

Watch as the applet repeats what you did: Shuffle the 30 cards and deal out 15 for the “Dolphin therapy” group, separating blue cards (successes) from green cards (failures), and 15 for the “Control Group”; create the table of “could have been” simulated results; add a dot to the dotplot for the number of improvers (blue cards) randomly assigned to the “Dolphin therapy” group.

Show Shuffle Options:

Number of Shuffles:

Cards  Data  Plot



(m) Now press **Shuffle** four more times. Does the number of successes vary among the repetitions?

(n) Enter **995** for the **Number of shuffles**. This produces a total of 1000 repetitions of the simulated random assignment process under the null hypothesis (the “null distribution”). What is the value of the mean of this null distribution? Explain why this center makes intuitive sense.

(o) Now we want to compare the observed result from the research study to these “could have been” results under the null hypothesis.

- How many improvers were there in the Dolphin therapy group in the actual study?
- Based on the resulting dotplot, does it seem like the actual experimental results would be surprising to arise solely from the random assignment process under the null model that dolphin therapy is not effective? Explain.

(p) In the **Count Samples** box, enter the observed statistic from the actual study and press **Count** to have the applet count the number of repetitions with 10 or more successes in the Dolphin group. In what proportion of your 1000 simulated random assignments were the results as (or more) extreme as the actual study?

(q) We said above that it would be equivalent to look at the difference in the conditional proportions. Use the **Statistic** pull-down menu (on far left side) to select **Difference in proportions**. How does the null distribution (including the mean) change? Does this make sense? How does the p-value change?

- (r) Interpret the p-value you found [*Hint:* Same reasoning as before, but now also focus on the source of randomness that was modelled in the simulation.]
- (s) Is this empirical p-value small enough to convince you that the experimental data that the researchers obtained provide strong evidence that dolphin therapy is effective (i.e., that the null model is not correct)?
- (t) Is it reasonable to attribute the observed difference in success proportions to the dolphin therapy alone? Explain.
- (u) To what population is it reasonable to generalize these results? Justify your answer.

**Discussion:** This procedure of randomly reassigning the response variable outcomes to the explanatory variable groups is often called a “randomization test.” The goal is to assess the chance variability from the random assignment process (as opposed to random sampling), though it is sometimes used to approximate the random chance arising from random sampling as well. Use of this procedure models the two-way table with fixed row and column totals (e.g., people were going to improve or not regardless of which treatment they received).

**Study Conclusions**

Due to the small p-value ( $0.01 \leq \text{p-value} \leq 0.05$ ) from our “randomization test,” we have moderately strong evidence that “luck of the draw” of the random assignment process alone is not a reasonable explanation for the higher proportion of subjects who substantially improved in the dolphin therapy group compared to the control group. Thus the researchers have moderate evidence that subjects aged 18–65 with a clinical diagnosis of mild to moderate depression (who would apply for such a study and fit the inclusion criteria) will have a higher probability of substantially reducing their depression symptoms if they are allowed to swim with dolphins rather than simply visiting and swimming in Honduras. Because this was a randomized comparative experiment, we can conclude that there is moderately strong evidence that swimming with dolphins *causes* a higher probability of experiencing substantial improvement in depression symptoms. But with this volunteer sample, we should be cautious about generalizing this conclusion beyond the population of people suffering from mild-to-moderate depression who could afford to take the time to travel to Honduras for two weeks.

**Practice Problem 3.6A**

A recent study compared two groups: One group was reminded of the sacrifices that physicians have to make as part of their training, and the other group was given no such reminder. All physicians in the study were then asked whether they consider it acceptable for physicians to receive free gifts from industry representatives. It turned out that 57/120 in the “sacrifice reminders” group answered that gifts are acceptable, compared to 13/60 in the “no reminder” group.

- (a) Identify the explanatory and response variables in this study.
- (b) Do you believe this study used random assignment? What would that involve?
- (c) Do you believe this study used random sampling? What would that involve?
- (d) Explain how you would carry out a simulation analysis to approximate a p-value for this study.  
*[Hint: How many cards? How many of each type? How many would you deal out? What would you record? How would you find the p-value?]*
- (e) A colleague points out that the sample sizes are not equal in this study, so we can't draw meaningful conclusions. How should you respond?

**Practice Problem 3.6B**

How would you find a two-sided p-value for the dolphin study?

### Investigation 3.7: Is Yawning Contagious?

The folks at *MythBusters*, a popular television program on the Discovery Channel, investigated whether yawning is contagious by recruiting fifty subjects at a local flea market and asking them to sit in one of three small rooms for a short period of time. For some of the subjects, the attendee yawned while leading them to the room (planting a yawn “seed”), whereas for other subjects the attendee did not yawn. As time passed, the researchers watched (via a hidden camera) to see which subjects yawned.

- (a) Identify the explanatory variable (EV) and the response variable (RV) in this study.

EV:

RV:

- (b) Define the relevant parameter of interest, and state the null and alternative hypotheses for this study. Be sure to clearly define any symbols that you use.

Parameter:

$H_0$ :

$H_a$ :

In the study they found that 10 of 34 subjects who had been given a yawn seed actually yawned themselves, compared with 4 of 16 subjects who had not been given a yawn seed.

- (c) More people yawned in the “seed” group (10 vs. 4), but this comparison is not meaningful by itself. Why not, and what would be a better comparison to make?

- (d) Create a two-table table summarizing the results, *using the explanatory variable as the column variable*.

		<b>Totals</b>
<b>Totals</b>		

- (e) Explain what will change about the simulation used to find an empirical p-value in this study as compared to the Dolphin Therapy study.

(f) Open the [Analyzing Two-way Tables](#) applet.

- Paste in the raw data and press **Use Data** or enter the titles and counts of a two-way table and press **Use Table**. (Or check the  $2 \times 2$  box and enter the cell values.)
- Check the **Show Shuffle Options** box.
- Set **Number of Shuffles** to **1000**.
- Press **Shuffle**.

		seed	no-seed
(explanatory,response)	yawn	10	4
	no-yawn	24	12

**Show Shuffle Options**

**Number of Shuffles**

**Shuffle**  **Data**  **Plot**

Briefly describe this randomization (null) distribution: What is its shape? What is the mean? What is the standard deviation?

(g) Specify the observed value for the difference in the conditional proportions in the **Count Samples** box. Then indicate whether the research conjecture expected a larger or smaller proportion of successes in Group A by choosing *Greater Than* or *Less Than* from the pull-down menu. Then press the **Count** button.

### Exact p-value

The simulations you have conducted in Investigations 3.6 (Dolphin Therapy) and above approximated the p-value for two-way tables arising from random assignment by assuming the row and column totals are fixed. In this case, the probability of obtaining a specific number of successes in one group can be calculated exactly using the *hypergeometric* probability distribution. (We used the independent *binomial* distributions with the teen hearing loss study, where we wanted to sample separately from two populations and the overall number of successes was not fixed in advance.)

Keep in mind, that under the null hypothesis, we are assuming the group assignments made no difference and that there would be 14 successes ("yawners") and 36 failures ("non-yawners") between the two groups regardless.

Because the random assignment makes every configuration of the subjects between the two groups equally likely, we determine the probability of any particular outcome for the number of yawners and non-yawners by first counting the total number of ways to assign 34 of the subjects to the yawn-seed group (and 16 to the no-yawn-seed group) in the denominator. The numerator is then the number of ways to get a particular set of configurations for that group, such as those consisting of 10 yawners and 24 non-yawners.

(h) How many ways altogether are there to randomly assign these 50 subjects into one group of 34 (yawn-seed group) and the remaining group of 16 (no-yawn-seed group)? [Hint: Recall what you saw earlier with the binomial distribution and counting the number of ways to obtain S successes and F failures in  $n$  trials. See the Technical Details in Investigation 1.1.]

(i) Now consider the 14 successes and the 36 failures. How many ways are there to randomly select 10 of the successes? How many ways are there to randomly assign 24 of the failures to be in the yawn seed group? How should you combine these two numbers to calculate the total number of ways to obtain 10 successes and 24 failures in the yawn-seed group, the configuration that we observed in the study?

<u>Successes</u>	<u>Failures</u>
------------------	-----------------

Total:

(j) To determine the exact probability that random assignment would produce exactly 10 successes and 24 failures into the group of 34 subjects, divide your calculation in (i) by your calculation in (h).

(k) Explain why your answer to (j) is *not* yet the p-value for this study.

**Result:** The probability of obtaining  $k$  successes in Group A, with  $n$  observations, when sampled from a two-way table with  $N$  observations, consisting of  $M$  successes and  $N-M$  failures is:

$$P(X = k) = C(M, k) \times C(N-M, n-k) / C(N, n)$$

where  $C(N, n) = N!/[n!(N-n)!]$  is the number of ways to choose  $n$  items from a group of  $N$  items.  $X$  represents the number of successes randomly selected for group A.  $X$  is a [hypergeometric](#) random variable. Also note  $E(X) = n(M/N)$  and  $SD(X) = \sqrt{nM(N-M)(N-n)/[N^2(N-1)]}$ .

In this study, we had  $N = 50$  subjects and we defined yawning to be success so  $M = 14$ . We also arbitrarily chose to focus on the yawn-seed group, so  $n = 34$ . This calculation works out the same if you had defined “not yawning” to be a success and/or if you had focused on the 16 people in the no-yawn-seed group. You just need to make sure you count consistently.

We will continue to define the p-value to be the probability of obtaining results *at least as extreme* as those observed in the actual study. Because we expected more yawners in the yawn-seed group, the p-value is the probability of randomly assigning *at least* 10 of the yawners in the yawn-seed group.

So far you have found  $P(X = 10) = C(14, 10) \times C(36, 24) / C(50, 34) = 0.2545$ .

- (l) Calculate  $P(X = 11)$ ,  $P(X = 12)$ ,  $P(X = 13)$ , and  $P(X = 14)$  using the hypergeometric probability formula.

$$P(X = 11)$$

$$P(X = 12)$$

$$P(X = 13)$$

$$P(X = 14)$$

Why do we stop at 14?

- (m) Sum all five probabilities together (including  $P(X = 10)$ ) to determine the exact p-value for the yawning study. How does this p-value compare to the empirical p-value from the applet simulation? Write a one or two sentence interpretation of this p-value.

Exact p-value:

Comparison:

Interpretation:

**Definition:** Using the hypergeometric probabilities to determine a p-value in this fashion for a two-way table is called [Fisher's Exact Test](#), named after R. A. Fisher.

- (n) Calculate this hypergeometric probability using technology (see Technology Detour on next page).

- (o) Set up and carry out the calculation to determine the exact p-value where you define the success to be “not yawning” and the group of interest to be the yawn seed group.

- (p) Set up and carry out the calculation to determine the exact p-value, where you focus on the number that did not yawn in the no-yawn-seed group. Show that you obtain the same exact p-value as before.

### Technology Detour – Calculating Hypergeometric Probabilities (Fisher’s Exact Test)

In R, the `iscamhyperprob` function takes the following inputs:

- $k$ , the observed value of interest (or the difference in conditional proportions, assumed if value is less than one, including negative)
- $total$ , the total number of observations in the two-way table
- $succ$ , the overall number of successes in the table
- $n$ , the number of observations in “group A”
- $lower.tail$ , a Boolean which is TRUE or FALSE

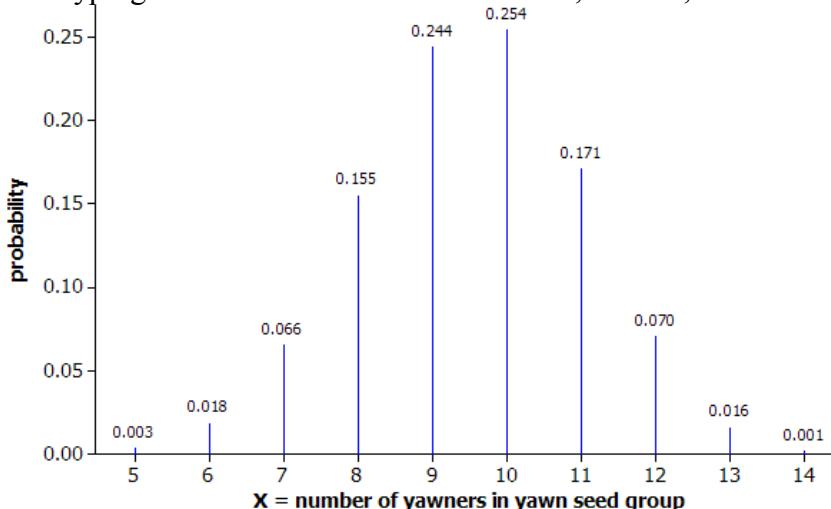
For example: `iscamhyperprob(k=10, total=50, succ=14, n=34, lower.tail=FALSE)`

#### Analyzing Two-way Tables applet

- Check the box for **Show Fisher’s Exact Test** in the lower left corner.
- A check box will appear for determining the two-sided p-value

**Discussion:** You should see that there are several equivalent ways to set up the probability calculation. Make sure it is clear how you define success/failure and which group you are considering “group A.” This will help you determine the numerical values for  $N$ ,  $M$ , and  $n$  in the calculation.

Below is a graph of the Hypergeometric distribution with  $N = 50$ ,  $M = 14$ , and  $n = 34$ .



Using probability rules, you can show that the expected value of this distribution is  $(M / N) \times n = (14/50) \times 34 = 9.52$  yawners in yawn seed group and the standard deviation of the probability distribution is the square root of  $n \times (M/N) \times (N-n)/N \times (N-M)/(N-1) = 1.496$  yawners.

(q) Compare this graph and the mean and standard deviation values to your simulation results.

(r) What conclusions will you draw from the p-value for this study?

- (s) On the *Mythbusters* program, the hosts concluded that, based on the observed difference in conditional proportions and the large sample size, there is “little doubt, yawning seems to be contagious.” Do you agree?

### Study Conclusions

With a large p-value of 0.513 (Fisher’s Exact Test), we do not have any evidence that the difference between the two groups (with and without yawn seed) was not created by chance alone from the random assignment process. If there was nothing to the theory that yawning is contagious, by “luck of the draw” alone, we would expect 10 or more of the yawners to end up in the yawn seed group in more than 50% of random assignments. Although the study results were in the conjectured direction, the difference between the yawning proportions was not large enough to convince us that the probability of yawning is truly larger when a yawn seed is planted. The researchers could try the study again with a larger sample size to increase the power of their test. The researchers also may want to be cautious in generalizing these results beyond the population of volunteers at a local flea market. It’s also not clear how naturalistic the setting of leading individuals to a small room to wait is.

**Practice Problem 3.7A**

- (a) For the Mythbusters' study ( $p\text{-value} > 0.5$ ), is it reasonable to conclude from this study that we have strong evidence that yawning is *not* contagious? Explain.
- (b) Explain, in this context, what is meant in the Study Conclusions box by "the researchers could try the study again with a larger sample size to increase the power of their test" and why that is a reasonable recommendation here.
- (c) To calculate the p-value here, why are we using the hypergeometric distribution instead of the binomial distribution?

**Practice Problem 3.7B**

Reconsider the Dolphin Therapy study (Investigation 3.6).

	<b>Dolphin Therapy</b>	<b>Control Group</b>	<b>Total</b>
<b>Showed substantial improvement</b>	10	3	13
<b>Did not show substantial improvement</b>	5	12	17
<b>Total</b>	15	15	30

Continue to focus on the number of improvers randomly assigned to the dolphin group, and represent this value by  $X$ .

- (a) When the null hypothesis is true, the random variable  $X$  has a hypergeometric distribution. Specify the values of  $N$ ,  $M$ , and  $n$ .
- (b) Calculating the exact p-value involves finding  $P(X \text{ _____})$ . [Hint: Fill in the blank with an inequality symbol and a number.]
- (c) Calculate this exact p-value, either by hand or with technology. Comment on whether this p-value is similar to the approximate one from your simulation results. (Be sure it's clear how you calculated this value.)
- (d) Suppose that the dolphin study had involved twice as many subjects, again with half randomly assigned to each group, and with the same proportion of improvers in each group. Determine the exact p-value in this case, and comment on whether/how it changes from the p-value with the real data. Explain why this makes sense.

### Investigation 3.8: CPR vs. Chest Compressions

For many years, if a person experienced a heart attack and a bystander called 911, the dispatcher instructed the bystander in how to administer chest compression plus mouth-to-mouth ventilation (a combination known as CPR) until the emergency team arrived. Some researchers believe that giving instruction in chest compression alone (CC) would be a more effective approach. In the 1990s, a randomized study was conducted in Seattle involving 518 cases (Hallstrom, Cobb, Johnson, & Copass, *New England Journal of Medicine*, 2000). In 278 cases, the dispatcher gave instructions in standard CPR to the bystander, and in the remaining 240 cases the dispatcher gave instructions in CC alone. A total of 64 patients survived to discharge from the hospital: 29 in the CPR group and 35 in the CC group.

- (a) Identify the observational units, explanatory variable, and response variable. Is this an observational study or an experiment?

Observational units:

Explanatory:

Response:

Type of study:      Observational      Experimental

- (b) Construct a two-way table to summarize the results of this study. Remember to put the explanatory variable in the columns.

- (c) Calculate the difference in the conditional proportions who survived (CC – CPR). Does this seem to be a noteworthy difference to you?

- (d) Use technology to carry out Fisher's Exact Test (by calculating the corresponding hypergeometric probability) to assess the strength of evidence that the probability of survival is higher with CC alone as compared to standard CPR. Write out how to calculate this probability, report the p-value, and interpret (on the next page) what it is the probability of.

$p\text{-value} = P(X \underline{\hspace{2cm}}) =$   
where  $X$  follows a hypergeometric distribution with  $N = \underline{\hspace{2cm}}$ ,  $M = \underline{\hspace{2cm}}$ , and  $n = \underline{\hspace{2cm}}$

Interpretation:

Because the sample sizes are large in this study, you should not be surprised that the distribution you examined in (d) is approximately normal. The large sample sizes allow us to approximate the hypergeometric distribution with a normal distribution. Thus, with large samples sizes (e.g., at least 5 successes and at least 5 failures in each group), an alternative to Fisher's Exact Test is the two-sample  $z$ -test that you studied in Section 3.1.

(e) Use technology to obtain the two-sample  $z$ -test statistic and p-value for this study. Compare this p-value to the one from Fisher's Exact Test; are they similar?

(f) Suggest a way of improving the approximation of the p-value.

(g) (Optional): Compare the normal approximation not the hypergeometric calculation.

[*Hints:* In R, use `iscamhypnorm(29, 518, 64, 278, TRUE)` or use the Analyzing Two-way Tables applet to compare the normal approximation to the hypergeometric calculation.]

(h) Do the data from this study provide convincing evidence that CC alone is better than standard CPR at the 10% significance level? Explain. How about the 5% level of significance?

(i) An advantage to using the  $z$ -procedures is being able to easily produce a confidence interval for the parameter. Use technology to determine a **90%** confidence interval for the parameter of interest, and then interpret this interval. [*Hint:* Think carefully about what the relevant parameter is in this study.]

(j) Suppose you had defined the parameter by subtracting in the other direction (e.g., CPR – CC instead of CC – CPR). How would that change:

(i) the observed statistic?

(ii) the test statistic?

(iii) the p-value?

(iv) confidence interval?

### Practice Problem 3.8

(a) Researchers in the CPR study also examined other response variables. For example, the 911 dispatcher's instructions were completely delivered in 62% of episodes assigned to chest compression plus mouth-to-mouth compared to 81% of the episodes assigned to chest compression alone.

(i) Calculate the difference in conditional proportions and compare it to the original study.

(ii) Without calculating, do you suspect the p-value for comparing this new response variable between the two groups will be larger or smaller or about the same as the p-value you determined above? Explain your reasoning.

(b) The above study was operationally identical to that of another study and the results of the two studies were combined. Of the 399 combined patients randomly assigned to standard CPR, 44 survived to discharge from the hospital. Of the 351 combined patients randomly assigned to chest compression alone, 47 survived to discharge.

(i) Calculate the difference in conditional proportions and compare it to the original study.

(ii) Without calculating, do you suspect the p-value for this comparison will be larger, smaller, or the same as the p-value you determined? Explain your reasoning.

## SECTION 4: OTHER STATISTICS

### Investigation 3.9: Flu Vaccine

A recent study (Jain et al, 2013) tested a vaccine for influenza in children ages 3 to 8. Children from 8 countries were randomly assigned to receive either the vaccine for influenza or a control (hepatitis A vaccine). An influenza-like illness was defined as a temperature of  $37.8^{\circ}\text{C}$  or higher, with at least one of the following: cough, sore throat, runny nose, or nasal congestion. Suspected cases were confirmed clinically. The results are shown in the table below:

	Hepatitis A vaccine ("control")	Quadrivalent influenza vaccine	Total
<b>Developed Influenza A or B</b>	148	62	210
<b>Did not develop influenza</b>	2436	2522	4958
<b>Total</b>	2584	2584	5168

- (a) Calculate the proportion of children developing influenza in each group. Does this appear to be a large difference to you?
- (b) Use Fisher's Exact Test to investigate whether these data provide convincing evidence that the probability of influenza is lower among children who receive the influenza vaccine. [Hint: State the hypotheses in symbols and in words. Define the random variable and outcomes of interest in computing your p-value.] Do you consider this strong evidence that the vaccine is effective?
- (c) Would you feel any differently about the magnitude of the difference in proportions if the conditional proportions with influenza had been 0.500 and 0.533? Explain.

**Discussion:** When the *baseline rate* (probability) of success is small, an alternative statistic to consider rather than the *difference* in the conditional proportions (which will also have to be small by the nature of the data) is the *ratio* of the conditional proportions. First used with medical studies where "success" is often defined to be an unpleasant event (e.g., death), this ratio was termed the *relative risk*.

**Definition:** The relative risk is the ratio of the conditional proportions, often intentionally set up so that the value is larger than one:

$$\text{Relative risk} = \frac{\text{Proportion of successes in group 1 (the larger proportion)}}{\text{Proportion of successes in group 2 (the smaller proportion)}}$$

The relative risk tells us how many times higher the “risk” or “likelihood” of “success” is in group 1 compared to group 2.

(d) Determine and interpret the ratio of the conditional proportions who developed influenza between the control and vaccinated groups in this study.

(e) Because we are now working with a ratio, we can also interpret this in terms of *percentage change*. Subtract one from the relative risk value and multiply by 100% to determine what percentage higher the proportion who developed influenza is in the control group compared to the vaccine group.

*Note:* The *efficacy* of a vaccine is defined as  $(1 - \text{unvaccinated risk} / \text{vaccinated risk}) \times 100\%$ .

Of course, now we would also like a confidence interval for the corresponding parameter, the ratio of the underlying probabilities of survival between these two treatments. When we produced confidence intervals for other parameters, we examined the sampling distribution of the corresponding statistic to see how values of that statistic varied under repeated random sampling. So now let's examine the behavior of the *relative risk* of conditional proportions using the [Analyzing Two-Way Tables](#) applet to simulate the *random assignment* process (as opposed to simulating the random sampling from a binomial process) under the (null) assumption that there's no difference between the two treatments. [See the Technology Detour below for equivalent R/Minitab instructions.]

(f) Generate a null distribution for Relative Risks:

- Check the **2x2** box
- Enter the two-way table into the applet and press **Use Table**.
- Generate 1000 random shuffles. This will take some time!
- Use the Statistic pull-down menu to select Relative Risk.

Describe the behavior of the null distribution of relative risk values.

(g) Where does the observed value of the relative risk from the actual study fall in the null distribution of the relative risks? What proportion of the simulated relative risks are at least this extreme?

(h) What percentage of the simulated relative risks are larger than 1.28 (just so you have a non-zero value to compare to later)?

But can we apply a mathematical model to this distribution?

- (i) Why does it make sense that the mean of the simulated relative risks is close to the value 1? [Hint: Remember the assumption behind your simulation analysis.]
- (j) You should notice some slight skewness in the distribution of relative risk values. Explain why you wouldn't be surprised for the distribution of this statistic to be skewed to the right (especially with smaller sample sizes).
- (k) In fact, this distribution is usually well modeled by a log normal distribution. To verify this, check the **ln relative risk** box (in the lower left corner) to take the natural log of each relative risk value and display a new histogram of these transformed values. Describe the shape of this distribution. Is the distribution of the *lnrelrisk* well modeled by a normal distribution?
- (l) What is the mean of the simulated *lnrelrisk* values? Why does this value make sense?
- (m) What is the standard deviation of the *lnrelrisk* values?
- (n) Calculate the observed value of  $\ln(\hat{p}_1 / \hat{p}_2)$  for this study. Where does this value fall (near in the middle or in the tail) of this simulated distribution of *lnrelrisk* values? Has the empirical p-value changed?
- (o) If you found the empirical p-value using  $\ln(\hat{p}_1 / \hat{p}_2)$ , it would be identical to the empirical p-value found in (i). Why? What did change about the distribution? [Hint: What percentage of the simulated *lnrelrisk* values are more extreme than  $\ln(1.28)$ , how does this compare to (h)?]

**Theoretical Result:** It can be shown that the standard error of the *ln relative risk* is approximated by  $SE\left(\ln \frac{\hat{p}_1}{\hat{p}_2}\right) = \sqrt{\frac{1}{A} - \frac{1}{A+C} + \frac{1}{B} - \frac{1}{B+D}}$

where  $A$ ,  $B$ ,  $C$ , and  $D$  are the observed counts in the  $2\times 2$  table of data, with  $A$  and  $B$  representing the number of “successes” in the two groups. Having this formula allows us to determine the variability from sample to sample without conducting the simulation first.

$A$	$B$
$C$	$D$
$A+C$	$B+D$

(p) Calculate the value of this standard error of the *ln(rel risk)* for this study. Interpret this value and compare it to the standard deviation from your simulated *lnrelrisk* values.

(q) You may find this approximate is in the ballpark but not all that close. What assumption is made by the simulation that is not made by this formula? What if you made the same assumption in this formula? [Hint: Think pooled  $\hat{p}$ .]

(r) Now that you have a statistic (*ln rel risk*) that has a sampling distribution that is approximately normal, what general formula can we use to determine a confidence interval for the parameter?

(s) Calculate the midpoint, 95% margin-of-error, and 95% confidence interval endpoints using the observed value of *ln(rel risk)* as the statistic and using the standard error calculated in (p).

(t) What parameter does the confidence interval in (s) estimate?

(u) Exponentiate the endpoints of this interval to obtain a confidence interval for the ratio of the probabilities of developing influenza between these two treatments. Interpret this interval.

(v) Is zero in this interval? Do we care? What value is of interest instead?

(w) Is the midpoint of this confidence interval for the population relative risk equal to the observed value of the sample relative risk? Explain why this makes sense.

(x) Suppose you used this method to construct a confidence interval for each of the 1,000 simulated random samples that you generated in (f). Do you expect the value 1 to be in these intervals? All of them? Most of them? What percentage of them? Explain.

(y) Compare the confidence interval you just calculated to the one given by the applet if you now check the **95% CI for relative risk** box.

### Study Conclusions

This study provided very strong evidence that children who receive the QIV vaccine are less likely to develop influenza than those who receive a control vaccine (exact one-sided p-value < 0.001,  $z$ -score = -6.08). An approximate 95% confidence interval for the difference in the probabilities indicates that the probability of influenza (of any severity) is 0.023 to 0.044 (2.3 to 4.4 percentage points) lower when receiving the QIV vaccine.

However, focusing on the difference in “success” probabilities has some limitations. In particular, if the probabilities are small it may be difficult for us to interpret the magnitude of the difference between the values. Also, we have to be very careful with our language, focusing on the difference in the influenza probabilities and not the percentage change. An alternative to examining a confidence interval for the difference in the conditional probabilities is to construct a confidence interval for the relative risk (ratio of conditional probabilities). A large sample approximation exists for a  $z$ -interval for the  $\ln(\text{relative risk})$  which can then be back-transformed to an interval for long-run relative risk. Many practitioners prefer focusing on this ratio parameter rather than the difference. From this study, we are 95% confident that ratio of the influenza probabilities is between 1.78 and 3.19. This means that receiving the placebo rather than the real vaccine raises the probability of developing influenza by between 78% and 219%.

An alternative would be to calculate the ratio in the other direction, using the relative risk of developing influenza between the vaccinated group compared to the unvaccinated group ( $\hat{p}_2 / \hat{p}_1 = 0.4189$ ). This approach is called examining the “efficacy” of the vaccination. Using the same standard error, a 95% confidence interval for the long-run relative risk is (0.313, 0.561). This means we are 95% confident that the use of the QIV vaccine reduces the probability of developing influenza by between  $(100 - 56.1) = 44\%$  and  $(100 - 31.3) = 69\%$ . In other words, we are 95% confident that the efficacy of the QIV vaccine in the larger population of children aged 3-8 years is between 44% and 69%.

**Note:** It can be risky to interpret the relative risk in isolation without considering the absolute risks (conditional proportions) as well. For example, doubling a very small probability may not be

noteworthy, depending on the context. You should also note that the percentage change calculation and interpretation depends on which group (e.g., treatment or control) is used as the reference group.

### Practice Problem 3.9

A multicenter, randomized, double-blind trial involved patients aged 36-65 years who had knee injuries consistent with a degenerative medial meniscus tear (Shivonen et al., *New England Journal of Medicine*, 2013). Patients received either the most common orthopedic procedure (arthroscopic partial meniscectomy,  $n_1 = 70$ ) or sham surgery that simulated the sounds, sensations, and timing of the real surgery ( $n_2 = 76$ ). After 12 months, 54 of those in the treatment group, reported satisfaction, compared to 53 in the sham surgery.

- Calculate and interpret a confidence interval for the ratio of the probabilities (relative risk) of satisfaction for these two procedures.
- What does your interval in (a) indicate about whether those receiving the orthopedic surgery are significantly more likely than those receiving a sham surgery to report satisfaction after 12 months? Explain your reasoning.

### Summary of Inference for “Relative Risk”

**Statistic:** ratio of conditional proportions (typically set up to be larger than one) =  $\hat{p}_1 / \hat{p}_2$

**Hypotheses:**  $H_0: \pi_1/\pi_2 = 1$ ;  $H_a: \pi_1/\pi_2 <, >, \text{ or } \neq 1$

**p-value:** Fisher’s Exact Test or normal approximation on  $\ln(\hat{p}_1 / \hat{p}_2)$

**Confidence interval for  $\pi_1/\pi_2$ :** exponentiate endpoints of  $\left[ \ln(\hat{p}_1 / \hat{p}_2) \pm z^* \sqrt{\frac{1}{A} - \frac{1}{A+C} + \frac{1}{B} - \frac{1}{B+D}} \right]$

**Note:** The confidence interval for the relative risk will not necessarily be symmetric around the sample statistic.

### Technology Detour – Simulating Random Assignment

We can select observations from a hypergeometric distribution for the cell 1 counts and then compute the cell 2 counts and the number of failures based on the fixed row and column totals. With this information you can compute the difference in conditional proportions, relative risk, etc. We should how to calculate  $\hat{p}_{\text{unvac}}$  below, the rest is up to you. Also keep in mind you can use “log” to calculate the natural logs of values. Also recall how you created a Boolean expression in Investigation 3.1 to find the p-value from the simulated results.

#### In R

```
> VacInfCount=rhyper(10000, 210, 4985, 2584)
  • 210 is the number of successes (M)
  • 4985 is the number of failures (N - M)
  • 2584 is the sample size (n)
> UnvacInfCount = 210-VacInfCount
> Unvacphat = UnvacInfCount/2584
```

### Investigation 3.10: Smoking and Lung Cancer

After World War II, evidence began mounting that there was a link between cigarette smoking and pulmonary carcinoma (lung cancer). In the 1950s, three now classic articles were published on the topic. One of these studies was conducted in the United States by Wynder and Graham (“Tobacco Smoking as a Possible Etiologic Factor in Bronchiogenic Cancer,” 1950, *Journal of the American Medical Association*). They found records from a large number of patients with a specific type of lung cancer in hospitals in California, Colorado, Missouri, New Jersey, New York, Ohio, Pennsylvania, and Utah. Of those in the study, the researchers focused on 605 male patients with this form of lung cancer. Another 780 male hospital patients with similar age and economic distributions without this type of lung cancer were interviewed in St. Louis, Boston, Cleveland, and Hines, IL. Subjects (or family members) were interviewed to assess their smoking habits, occupation, education, etc. The table below classifies them as non-smoker or light smoker, or at least a moderate smoker.

**Wynder and Graham**

	<b>None or Light smoker (0-9 per day)</b>	<b>Moderate to Heavy smoker (10-35+ per day)</b>	<b>Total</b>
<b>Lung cancer patients</b>	22	583	605
<b>Controls</b>	204	576	780
<b>Total</b>	226	1159	1385

(a) Calculate and interpret the relative risk of being a lung cancer patient for the moderate to heavy (“regular”) smokers compared to the None or Light “non-smokers.”

(b) Does this feel like an impressive statistic to you? Do you think it will be statistically significant?

(c) What is the estimate of the baseline rate of lung cancer from this table? Does that seem to be a reasonable estimate to you? How is this related to the design of the study?

(d) Calculate and interpret the relative risk of being a control patient for the non-smokers compared to the regular smokers. How does this compare to (a) and (b)?

**Definition:** There are three main types of observational studies.

- *Cross-classification study.* The researchers categorize subjects according to both the explanatory and the response variable simultaneously. For example, they could take a sample of adult males and simultaneously record both their smoking status and whether they have lung cancer. A common design is *cross-sectional*, where all observations are taken at a fixed point in time.
- *Cohort study.* The researchers identify individuals according to the explanatory variable and then observe the outcomes of the response variable. These are usually *prospective designs* and may even follow the subjects (the *cohort*) for several years.
- *Case-control study.* The researchers identify observational units in each response variable category (the “cases” and the “controls”) and then determine the explanatory variable outcome for each observational unit. How the controls are selected is very important in determining the comparability of the groups. These are often *retrospective designs* in that the researchers may need to “look back” at historical data on the observational units.

(e) Would you classify the Wynder & Graham study as cross-classified, cohort, or case-control? Explain.

(f) Explain why using the relative risk (or even the difference in proportions) as the statistic can be problematic with case-control studies.

An advantage of case-control studies is when you are studying a “rare event,” you can ensure a large enough number of “successes” and fairly balanced group sizes. However, a disadvantage is that it does not make sense to calculate “risk” or likelihood of success from a case-control study, because the distribution of the response variable has been manipulated/determined by the researcher. Switching the roles of the explanatory and response often gives very different results for relative risk (changing our measure of the strength of the relationship) and often really isn’t the comparison of interest stated by the research question. Consequently, **conditional proportions of success and relative risk are not appropriate statistics to use with case-control studies**. Instead, we will consider another way to compare the uncertainty of an outcome between two groups.

**Definition:** The odds of success are defined as the ratio of the proportion of “successes” to the proportion of “failures,” which simplifies to the ratio of the number of successes to failures.

$$\text{odds} = \frac{\text{proportion of successes in the group}}{\text{proportion of failures in the group}} = \frac{\text{number of successes in the group}}{\text{number of failures in the group}}$$

For example, if the odds are 2-to-1 in favor of an outcome, we expect a success twice as often as a failure in the long run, so this corresponds to a probability of 2/3 of the outcome occurring. Similarly, if the probability of success is 1/10, then the odds equals  $(1/10)/(9/10) = 1/9$ , and failures is 9 times more likely than success. It’s important to note how the “outcome” is defined. For example, in horse racing, odds are typically presented in terms of “losing the race,” so if a horse is given 2-to-1 odds against winning a race, we expect the horse to lose two-thirds of the races in the long run.

**Definition:** The *odds ratio* is another way to compare conditional proportions in a  $2 \times 2$  table.

$$\text{Odds ratio} = \frac{\text{(number of successes in group 1/number of failures in group 1)}}{\text{(number of successes in group 2/number of failures in group 2)}}$$

Like relative risk, if the odds ratio is 3, this is interpreted as “the odds of success in the ‘top’ group are 3 times (or 200%) higher than the odds of success in the ‘bottom’ group.” However, the relative risk and the odds ratio are not always similar in value.

(g) Calculate and interpret the odds ratio comparing the odds of lung cancer for the smokers to the odds of lung cancer for the control group. Does this match (a)?

(h) Calculate and interpret the odds ratio for being in the control group for the non-smokers compared to the smokers? Does this match (d) or (g)?

**Key Results:** A major disadvantage to relative risk is that your (descriptive) measure of the strength of evidence that one group is “better” depends on which outcome you define a success as well as which variable you treat as the explanatory and which as the response. But a big advantage to odds ratio is that it is *invariant* to these definitions (If your odds are 10 times higher to die from lung cancer if you are a smoker, then your odds of being a smoker are 10 times higher if you died from lung cancer). The only real disadvantage is that the odds ratio is trickier to interpret (“higher odds” vs. the more natural “more likely”). Thus, for case-control studies in particular, the odds ratio is the preferred statistic. However, when the success proportions are both small, the odds ratio can be used to approximate the relative risk.

(i) Let  $\tau$  (“tau”) represent the population odds ratio of having lung cancer for those who are regular smokers compared to those who are not regular smokers, so  $\tau = \pi_1/(1 - \pi_1)/(\pi_2/(1 - \pi_2))$ . State the null and alternative hypotheses in terms of this parameter.

(j) Use Fisher’s Exact Test to calculate the p-value. (Note: We get the same p-value no matter which statistic we use, why is that?)

**Theoretical Result:** The sampling distribution of the sample odds ratio also follow a log-normal distribution like the relative risk (for any study design). Thus, we can construct a confidence interval for the population/treatment log-odds ratio using the normal distribution. The standard error of the sample log-odds ratio (using the natural log) is given by the expression:

$$SE(\ln \text{ odds ratio}) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \text{ where } A, B, C, \text{ and } D \text{ are the four table counts.}$$

(k) Calculate this standard error and then use it to find an approximate 95% confidence interval for this odds ratio.

(l) Back-transform the end-points of the interval and (k) and interpret your results.

(m) Does your interval contain the value one? Discuss the implications of whether or not the interval contains the value one.

(n) Compare your results to the following R output:

```
> fisher.test(matrix(c(583, 576, 22, 204), nrow=2), alt = "two.sided")
Fisher's Exact Test for Count Data

data: matrix(c(583, 576, 22, 204), nrow = 2)
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 5.918014 15.519781
sample estimates:
odds ratio
 9.372535
```

(o) Summarize (with justification) the conclusions you would draw from this study (using both the p-value and the confidence interval, and addressing both the population you are willing to generalize to and whether or not you are drawing a cause-and-effect conclusion).

## Study Conclusions

Because the baseline incidence of lung cancer in the population is so small, the researchers conducted a *case-control* study to ensure they would have both patients with and without lung cancer in their study (matched by age and economic status). In a case-control study, the *odds ratio* is a more meaningful statistic to compare the incidence of lung cancer between the two groups. We find that the sample odds of lung cancer are almost ten times larger for the regular smokers compared to the non-regulars in this study. By the *invariance* of the odds ratio, this also tells us that the odds of being a regular smoker (rather than not) are almost 10 times higher for those with lung cancer. We are 95% confident that in the larger populations represented by these samples, the odds of lung cancer are 5.92 to 15.52 times larger for the regular smokers (Fisher's Exact Test p-value << 0.001). If both success proportions had been small, we could say this is approximately equal to the relative risk and use the words "10 times higher" or "10 times more likely." The full data set (which broke down the second category further) also shows that the odds of having lung cancer increase with the amount of smoking (light smokers have 2 times the odds, heavy smokers have 11 times the odds, and chain smokers have 29 times the odds!) – this is called a "dose-response." We see a strong relationship between the size of the "dose" of smoking and occurrence of lung cancer for these patients.

However, this study was criticized for "retrospective bias" in asking subjects to accurately remember, and be willing to tell, details of their lifestyles. This can also be complicated by asking these questions of patients who know they have been diagnosed with lung cancer, as their recall may be affected by this knowledge. We also have to worry whether hospitalized males are representative of the male population.

Other studies around the same time (e.g., Hammond and Horn, Wynder and Cornfield) found similar increases in "risk" with smoking. However, these were all observational studies so critics reasonably argued that other variables such as lifestyle, diet, exercise, and genetics could be responsible for both the smoking habits and the development of lung cancer. Although there was still much (on-going) research to be done, and these studies did not claim to *prove* that cigarette smoking causes lung cancer, these landmark studies set the stage. They also led to many efforts in improving study design and in developing statistical tools (such as relative risk and odds ratios) to analyze the results.

## Practice Problem 3.10A

A researcher searched court records to find 908 individuals who had been victims of abuse as children (11 years or younger). She then found 667 individuals, with similar demographic characteristics, who had not been abused as children. Based on a search through subsequent years of court records, she determined how many in each of these groups became involved in violent crimes (Widom, 1989). The results are shown below:

	Abuse victim	Control
Involved in violent crime	102	53
Not involved in violent crime	806	614

- (a) Is this an observational study or an experiment? If observational, which type?
- (b) Calculate and interpret the odds ratio of being involved in a violent crime between these two groups.
- (c) The one-sided p-value for this result (using Fisher's Exact Test) is 0.018 (confirm). Is it reasonable to conclude that being a victim of abuse as a child causes individuals to be more likely to be violent toward others afterwards? Explain.
- (d) Calculate and interpret a 95% confidence interval for the population odds ratio.
- (e) Is it reasonable to generalize these results to all abuse and non-abuse victims? Explain.

**Practice Problem 3.10B**

- (a) Suppose that individuals in Group 1 have a 2/3 probability of success, and those in Group 2 have a 1/2 probability of success. Calculate and interpret the relative risk of success, comparing Group 1 to Group 2.
- (b) Calculate and interpret the odds of success for Group 1.
- (c) Calculate and interpret the odds ratio of success, comparing Group 1 to Group 2.
- (d) Suppose Group 3 has a 0.1 probability of success, and Group 4 has a 0.05 probability of success. Repeat questions (a) and (c).
- (e) In which case (Groups 1 and 2, or Groups 3 and 4) are the relative risk and odds ratio more similar? Why?

**Summary of Inference for Odds Ratio**

*Statistic:*  $\hat{\tau} = [\hat{p}_1/(1 - \hat{p}_1)] / [\hat{p}_2/(1 - \hat{p}_2)] = (A \times D) / (B \times C)$   
 (typically set up to be larger than one)

A	B
C	D

*Hypotheses:*  $H_0: \tau = 1; H_a: \tau <, >, \text{ or } \neq 1$

*p-value:* Fisher's Exact Test or normal approximation on  $\ln(\hat{\tau})$

*confidence interval for  $\tau$ :* exponential of  $\left[ \ln(\hat{\tau}) \pm z^* \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \right]$

**In R:** > fisher.test(matrix(c(a, c, b, d), nrow=2), alt = )

### Investigation 3.11: Sleepy Drivers

Connor et al. (*British Medical Journal*, May 2002) reported on a study that investigated whether sleeplessness is related to car crashes. The researchers identified all drivers or passengers of eligible light vehicles who were admitted to a hospital or died as a result of a car crash on public roads in the Auckland, New Zealand region between April 1998 and July 1999. Though cluster sampling, they identified a sample of 571 drivers who had been involved in a crash resulting in injury and a sample of 588 drivers who had not been involved in such a crash as representative of people driving on the region's roads during the study period. The researchers asked the individuals (or proxy interviewees) whether they had a full night's sleep (at least seven hours mostly between 11pm and 7am) any night during the previous week. The researchers found that 61 of the 535 crash drivers who responded and 44 of the 588 "no crash" drivers had not gotten at least one full night's sleep in the previous week.

- (a) Identify the observational units and variables in this study. Which variable would you consider the explanatory variable and which the response variable? Was this an observational study or an experiment? If observational, would it be considered a case-control, cohort, or cross-classified design?

Observational units:

Explanatory variable:

Response variable:

Type of study:

- (b) Organize these sample data into a  $2 \times 2$  table:

	No full night's sleep in past week ("sleep deprived")	At least one full night's sleep in past week ("not sleep deprived")	Sample sizes
Crash			535
No crash			588
Total			1123

- (c) Which statistic (odds ratio or relative risk) is most appropriate to calculate from this table, considering how the data were collected? Calculate and interpret this statistic. Does the value of this statistic support the researchers' conjecture? Explain.

**Statistical Inference**

(d) Outline the steps of a simulation that models the randomness in this study and helps you assess how unusual the statistic is that you calculated in (c) when the null hypothesis is true. Include a statement of the null and alternative hypotheses for your choice of parameter.

(e) Use technology to carry out your simulation and draw your conclusions. [*Hint:* Be careful of rounding issues in finding your p-value, make sure you are including observations as extreme as the observed in your count.]

(f) Calculate and interpret a 95% confidence interval for your choice of parameter.

(g) Summarize (with justification) the conclusions you would draw from this study (using both the p-value and the confidence interval, and addressing both the population you are willing to generalize to and whether or not you are drawing a cause-and-effect conclusion).

## Study Conclusions

The proportions of drivers who had not gotten a full night's sleep in the previous week were 0.107 for the case group of drivers who had been involved in a crash, compared to 0.075 for the control group who had not. Because these proportions are small, and because of the awkward roles of the explanatory and response variables in this study (we would much rather make a statement about the proportion of sleepless drivers who are involved in crashes), the odds ratio is a more meaningful statistic to calculate. The sample odds of having missed out on a full night's sleep were 1.59 times higher for the case group than for the control group. By the invariance of the odds ratio, we can also state that the sample odds of having an accident are 1.59 times (almost 60%) higher for those who do not get a full night sleep than those who do. The empirical p-value (less than 5%) provides moderately strong evidence that such an extreme value for the sample odds ratio is unlikely to have arisen by chance alone if the proportion of drivers with sleepless nights was 0.09 for both the population of "cases" and the population of "controls." (Using a one-sided Fisher's Exact Test, we get p-value = 0.016.) A 95% confidence interval for the population odds ratio extends from 1.06 to 2.39 (1.04 to 2.45 with R). This interval provides statistically significant evidence that the population odds ratio exceeds one and that, with 95% confidence, the odds of having an accident are about 1 to 2.5 times higher for the sleepy drivers than for well rested drivers. We cannot attribute this association to a cause-and-effect relationship because this was an observational (case-control) study. We might also want to restrict our conclusions to New Zealand drivers.

## Practice Problem 3.11

Another landmark study on smoking began in 1952 (Hammond and Horn, 1958, "Smoking and death rates—Report on forty-four months of follow-up of 187,783 men: II. Death rates by cause," *JAMA*). They used 22,000 American Cancer Society volunteers as interviewers. Each interviewer was to ask 10 healthy white men between the ages of 50 and 69 to complete a questionnaire on smoking habits. Each year during the 44-month follow-up, the interviewer reported whether or not the man had died, and if so, how. They ended up tracking 187,783 men in nine states (CA, IL, IA, MI, MN, NJ, NY, PA, WI). Almost 188,000 were followed up by the volunteers through October 1955, during which time about 11,870 of the men had died, 488 from lung cancer. The following table classifies the men as *having a history of regular cigarette smoking or not and whether or not they died from lung cancer*. In this study, nonsmokers are grouped with occasional smokers, including pipe- and cigar-only smokers.

### Hammond and Horn

	Not regular smoker	Regular smoker	Total
Lung cancer death	51	397	448
Alive or other cause of death	108,778	78,557	187,335
Total	108,829	78,954	187,783

- (a) Is this a case-control, cohort, or cross-classified study?
- (b) Calculate and interpret an odds ratio from the two-way table.
- (c) Produce and interpret a 95% confidence interval for the population odds ratio.
- (d) Are these results consistent with the Wynder and Graham study? Explain.

**Example 3.1: Wording of Questions**

*Try these questions yourself before you use the solutions following to check your answers.*

Researchers have conjectured that the use of the words “forbid” and “allow” can affect people’s responses to survey questions. Students in an introductory statistics class were randomly assigned to answer one of the following questions:

- Should your college allow speeches on campus that might incite violence?
- Should your college forbid speeches on campus that might incite violence?

Of the 11 students who received the first question, 8 responded yes. Of the 14 students who received the second question, 12 said no.

(a) Identify the observational units and the explanatory and response variables.

(b) Is this an observational study or an experiment? If an observational study, suggest a potential confounding variable. If an experiment, explain the roles of randomization and blinding in this study.

(c) Construct a two-way table to summarize these results.

(d) Construct a segmented bar graph to display these results and comment on the relationship revealed by this graph.

(e) Based on earlier studies, researchers expected people to be less likely to agree to “forbid” the speeches, leading to more no responses (and thus appearing to be in favor of having the speeches), whereas they expected people to be comparatively less likely to agree to “allow” the speeches. Do these data provide strong evidence that these students responded more positively toward having such speeches

if their question was phrased in terms of “forbid” rather than “allow”? Carry out a test of significance and explain the decision you would make based on the p-value. Write a paragraph summarizing your conclusions including whether a cause-and-effect conclusion can be drawn and the population you are willing to generalize these results to.

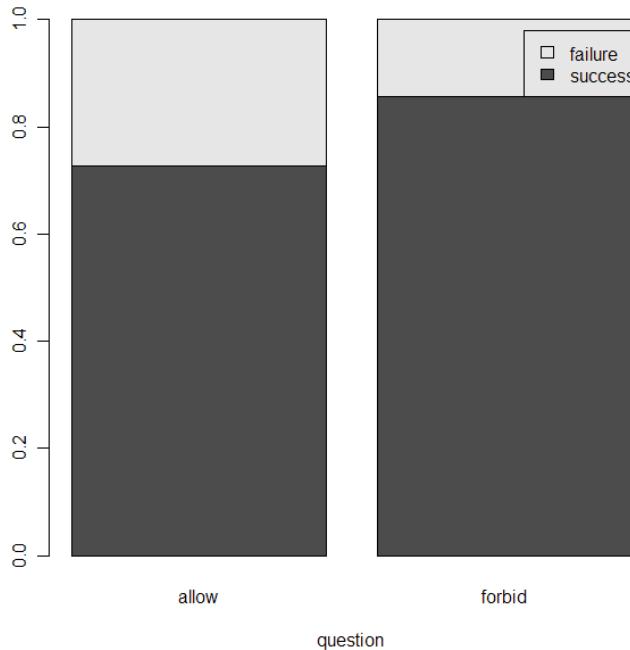
- (f) In a 1976 study, one group of subjects was asked, “Do you think the United States should forbid public speeches in favor of communism?”, whereas another group was asked, “Do you think the United States should allow public speeches in favor of communism?”. Of the 409 subjects randomly asked the “forbid” version, 161 favored the forbidding of communist speeches. Of the 432 subjects asked the “allow” version, 189 favored allowing the speeches. Construct a segmented bar graph for these data and comment on whether you believe the p-value for this table will be larger or smaller than that in (e). Explain your reasoning.

## Analysis

- (a) The observational units in this study are the statistics students. The explanatory variable is the word choice in the question they responded to (categorical) and the response variable is whether their response was in favor of the speeches (“yes” with the allow question and “no” with the forbid question, categorical).
- (b) This was an experiment because the word choice in the question was randomly assigned to the students. Presumably the instructor mixed up the order of the questionnaires prior to passing them out to the students. This is important to equalize other variables between these two groups such as political inclinations and gender. The students did not know that there were two different forms of the questions, so the study was blind. If they had realized that the instructor was focusing on how they responded to the two words, they probably would have responded differently eliminating any subconscious effect of the word choice.
- (c) Two-way table, with the explanatory variable, word choice, as the column variable, and defining a “success” to mean that the student is in favor of such speeches:

	Allow	Forbid	Total
Success	8	12	20
Failures	3	2	5
Total	11	14	25

(d) Here is the segmented bar graph, with word choice along the horizontal axis, and success defined as favoring the speeches (yes to the allow question and no to the forbid question).



We see that most of these students were in favor of the speeches (80%). There was a slight tendency for those responding to the forbid question to appear more in favor (more likely to say no, do not forbid the speeches), 0.857 versus 0.727. However, the distribution within the bars look fairly similar and the association does not appear to be strong.

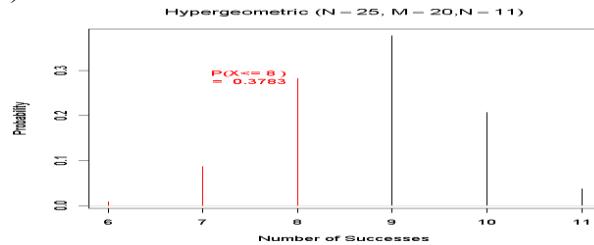
(e) The null hypothesis will be that there is no effect due to wording of the question. So if  $\pi_{\text{allow}}$  is the probability someone says “yes” to the allow question and if  $\pi_{\text{forbid}}$  is the probability someone says no to the forbid question:

$$H_0: \pi_{\text{allow}} - \pi_{\text{forbid}} = 0$$

and based on the prior research:  $H_a: \pi_{\text{allow}} - \pi_{\text{forbid}} < 0$

Fisher’s Exact Test indicates how often we expect to see as few as 8 or fewer successes in the allow group (equivalently, at least as many as 12 successes in the forbid group). So if we define X to be the number in the allow group in favor of the speeches, X follows a hypergeometric distribution with  $N = 25$ ,  $M = 20$ , and  $n = 11$ . The p-value will be  $P(X \leq 8)$ .

$$P(X \leq 8) = \frac{\binom{20}{8} \binom{5}{3}}{\binom{25}{11}} + \frac{\binom{20}{7} \binom{5}{4}}{\binom{25}{11}} + \frac{\binom{20}{6} \binom{5}{5}}{\binom{25}{11}} \\ = 0.3783$$



Confirming our calculations using R:

Fisher's Exact Test for Count Data

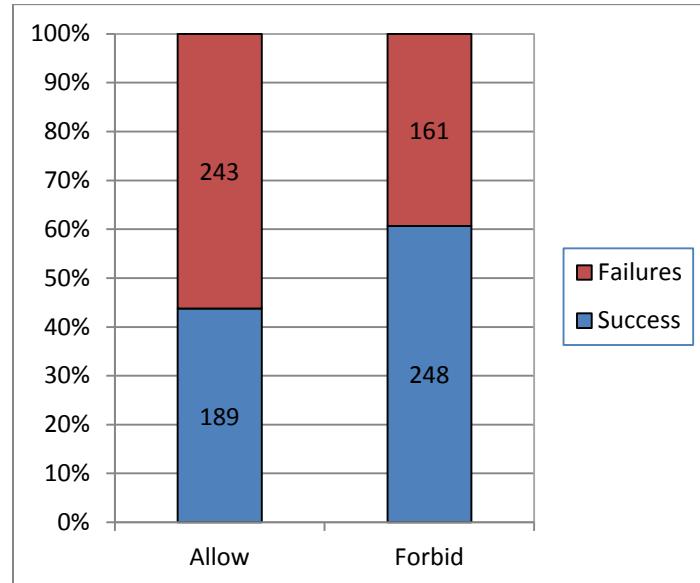
```
data: table
p-value = 0.3783
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
0.000000 3.611826
sample estimates:
odds ratio
0.4594844
```

Note, it would NOT be valid to use the normal approximation here because we do not have at least 5 failures in both groups (normal approx. p-value equals 0.2149, with continuity correction equals 0.384)

Thus, if the word choice in the question had no effect, we would get experimental results at least this extreme in about 38% of random assignments. This indicates that our experimental data is not surprising and does not provide evidence that the wording of the question had an effect on these students. Thus, we will not say the word choice in the question made a difference (though if the p-value had been small, because this was a randomized experiment, a cause-and-effect conclusion would have been valid). Furthermore, we should be hesitant in generalizing these results beyond introductory statistics students at this school. We do not know how these students were selected from this school nor whether these students might be representative of other college students. Perhaps this is a private school or the students tend to be more liberal than at other schools.

(f) If we again look at a segmented bar graph:

We see that in this study, individuals were less likely to be in favor of the speeches in general (52%). More importantly, the difference between the two groups is a bit larger ( $0.606 - 0.438 = 0.169$  in this study, compared to a difference of 0.130 before), providing stronger evidence that such a difference would not happen due to chance alone and thus a smaller p-value.



We must also take into account the fact that the group sizes are much, much larger in this study. With such large sample sizes, even small differences between the groups may appear statistically significant. Thus, with the larger samples and the slightly larger difference in success proportions, the p-value from this study will be much smaller than that from the original study.

**Example 3.2: Worries about Terrorist Attacks**

*Try these questions yourself before you use the solutions following to check your answers.*

A 2004 Harris Interactive poll asked respondents “How often do you worry about the possibility of a terrorist attack in this country?” Parallel surveys were conducted online of a “nationwide cross-section” of adults over the age 18 in the United States (between January 19 and 28) and in Great Britain (between January 21 and 26). Of 3,378 American respondents, 45% don’t worry much or at all compared to 41% of 2,417 British respondents.

(a) Carry out a test of significance to determine whether the difference in these sample proportions is convincing evidence of a difference in the population proportions. What conclusion does the test allow you to draw about the population proportions who don’t worry much or at all about terrorist attacks?

(b) The responses to two other survey questions are given below. Without calculating, would the p-values for each of these comparisons be larger or smaller to the p-value in (a)?

	US	UK
"How likely do you think it is that there will be a major terrorist attack in this country in the next twelve months?"	62% thought likely	64% thought likely
"How much confidence do you have in the ability of the government to reduce the likelihood of a terrorist attack?"	72% greatly or somewhat confident	54% greatly or somewhat confident

(c) The Harris Interactive website included the following statement:

In theory, with probability samples of this size, one could say with 95 percent certainty that the results have a statistical precision of plus or minus 1.7 percentage points (in the US poll, 2 percentage points in the UK poll) of what they would be if the entire adult population had been polled with complete accuracy. Unfortunately, there are several other possible sources of error in all polls or surveys that are probably more serious than theoretical calculations of sampling error. They include refusals to be interviewed (non-response), question wording and question order, and weighting. It is impossible to quantify the errors that may result from these factors. This online sample is not a probability sample.

- Where do the “1.7 percentage points” and “2 percentage points” figures come from?
  
  
  
- Why do you think they state that this online sample is not a probability sample?

## Analysis

(a) The response variable is categorical and binary. Suppose we define “success” as “not worrying much or at all about a terrorist attack in their country.” Then we can let  $\pi_{\text{US}} - \pi_{\text{UK}}$  represent the difference in the population proportions that would respond that they don’t worry much or at all.

$$H_0: \pi_{\text{US}} - \pi_{\text{UK}} = 0 \text{ (there is no difference in the population proportions)}$$

$$H_a: \pi_{\text{US}} - \pi_{\text{UK}} \neq 0 \text{ (there is a difference)}$$

Note: We were not given a conjectured direction as to which country would have a higher population proportion.

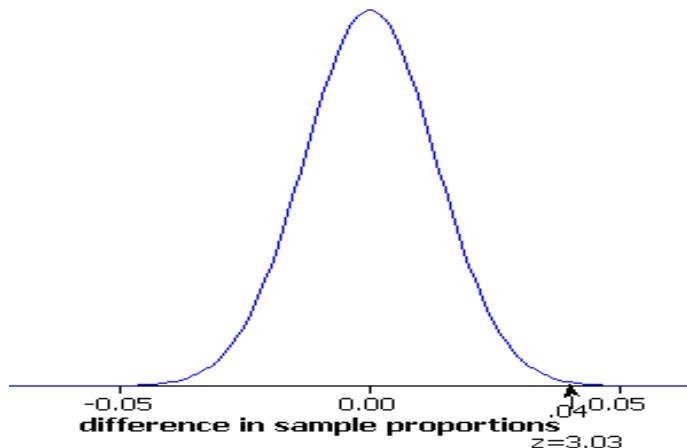
Because these polls were conducted separately in the two countries, the samples are independent. However, as discussed in (c), they are not true random samples, so we need to have some caution in generalizing these results to all adults over the age of 18 in the two countries.

The sample sizes were large (over 900 successes and failures in each sample) so the normal distribution would be a reasonable model for the sampling distribution of the difference in sample proportions.

Using this normal distribution model, we would expect the distribution of the difference in sample proportions to be centered at zero. Assuming that the null hypothesis of equal population proportions is true, our estimate of the standard deviation of this distribution is

$$\sqrt{0.433(1-0.433)\left(\frac{1}{3378} + \frac{1}{2417}\right)} = 0.0132$$

where  $0.433 = (1520+991)/(3378+2417)$ , the pooled estimate of the proportion of successes.

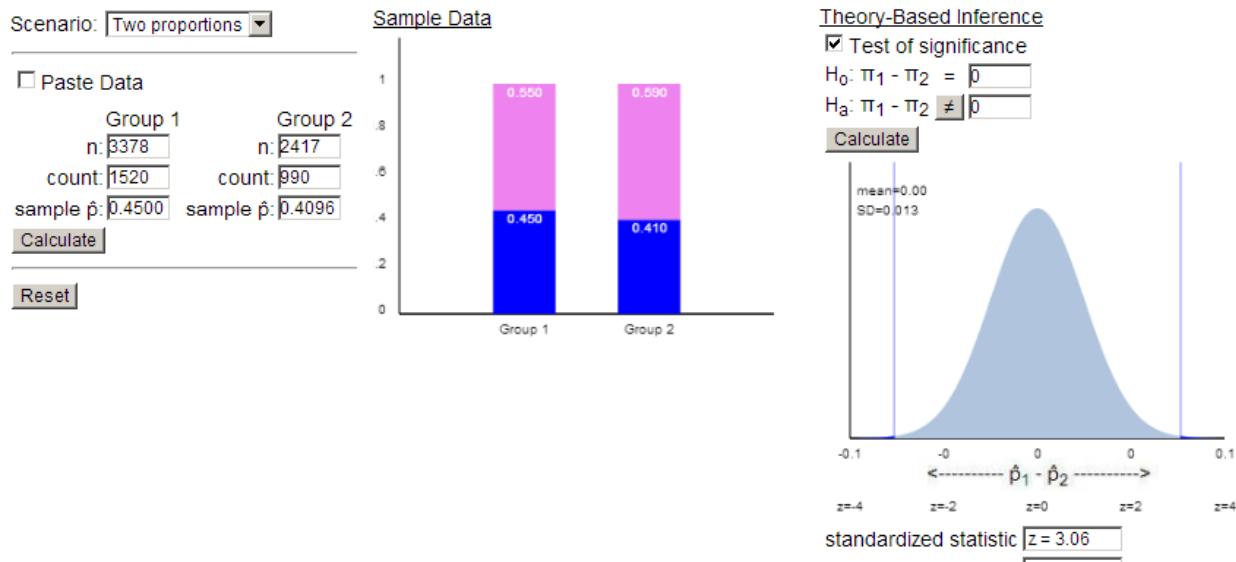


With this standard error, the observed difference in the sample proportions ( $.45 - .41 = 0.04$ ) is 3.03 standard errors from the hypothesized difference of zero:

$$z_0 = \frac{0.45 - 0.41}{0.0132} = 3.03$$

The probability of observing a test statistic at least this extreme (in either direction) by random sampling alone is p-value =  $2P(Z \geq 3.03) = 2(0.0012) = 0.0024$ .

These calculations (apart from rounding discrepancies) are confirmed by R and the Theory-Based Inference applet and Minitab:



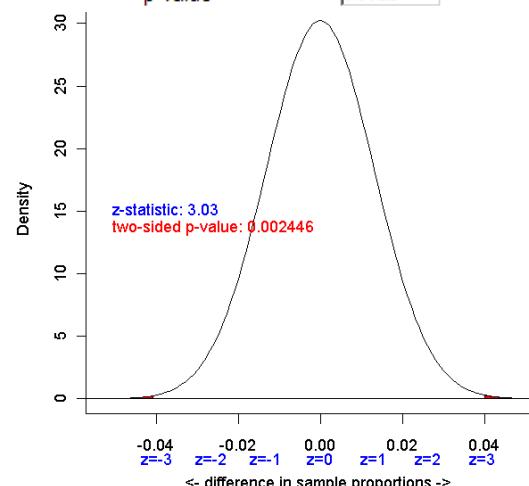
### Two Proportion z test

Group1: observed successes = 1520,  
sample size = 3378, sample proportion = 0.45

Group2: observed successes = 991, sample size = 2417, sample proportion = 0.41

Null hypothesis :  $\pi_1 - \pi_2 = 0$   
Alternative hypothesis:  $\pi_1 - \pi_2 \neq 0$   
z-statistic: 3.03  
p-value: 0.002446

95 % Confidence interval for  $\pi_1 - \pi_2$ :  
( 0.01419456 , 0.06580544 )



### Test and CI for Two Proportions

Sample	X	N	Sample p
1	1520	3378	0.449970
2	990	2417	0.409599

Difference = p (1) - p (2)  
Estimate for difference: 0.0403717  
95% CI for difference: (0.0145686, 0.0661749)  
Test for difference = 0 (vs not = 0): Z = 3.06 P-Value = 0.002

Fisher's exact test: P-Value = 0.002

With such a small p-value (e.g., less than 0.01), we reject the null hypothesis. If the population proportions had been equal, this tells us that there is a very small probability of random sampling alone leading to sample proportions at least as far apart as those found by Harris. Therefore, we have strong

evidence that the population proportions who worry little or not at all about terrorist attack is not the same in the US and UK.

A 95% confidence interval for the difference in the population proportions is (0.014, 0.066). So we are 95% confident that percentage of all Americans that were not worried was (in 2004) 1.4 to 6.6 percentage points higher than all Britons that were not worried about terrorist attack. This is not a large difference in a practical sense, but we don't believe a difference in sample proportions this large could have arisen by random sampling alone (in particular due to the large sample sizes).

(b) Because the sample sizes are the same but the difference in sample proportions would be smaller, the p-value for the “likelihood of attack” question would be larger. On the other hand, because the sample proportions are further apart for the “confidence in government to prevent attacks” question, the p-value would be even smaller. Thus, there would be more evidence that the populations of Americans and Britons differ in their opinions about government control over terrorism, but less evidence that they differ in their expectations for an attack.

(c) If we assume a population proportion of around 0.5, then for the US sample, the margin-of-error for a 95% confidence interval would be:

$$1.96 \sqrt{0.5(0.5)/3378} = 0.0169.$$

The margin-of-error for the UK sample will be a bit larger because the sample size was smaller:

$$1.96 \sqrt{0.5(0.5)/2417} = 0.0199.$$

If the population proportion turns out to be different from 0.5, the actual margin-of-error would be smaller. Thus, these calculations give us quick approximations to the maximal margins-of-error.

The above interpretations (p-value, confidence level, margin-of-error) rely on the samples being randomly selected from each population, so Harris provides a warning that their sampling method is not by definition a probability sample. Because of this, we must have some cautions in generalizing these results to the larger populations (though the methods were at least similar in the two countries). Still, further investigation of Harris' methodology (e.g., <http://www.harrisinteractive.com/MethodsTools/DataCollection.aspx>) reveals that they use many methods to “maintain the reliability and integrity in the sample” and the web is used as the response mechanism instead of, say, the telephone. This makes the Harris online poll much better than the ubiquitous “click here” non-scientific opinion polls found all over the web. So although we must always worry about non-sampling errors (Investigation 1.15) with any poll, these Harris polls can be considered as reliable as a telephone poll.

## CHAPTER 3 SUMMARY

In this chapter, you focused on comparing two groups on a categorical variable. You began by extending the simulations of sampling from a finite population to consider two independent random samples from the same population (same probability of success). You used a normal approximation to again find approximate p-values and confidence intervals. You were soon cautioned against jumping to cause-and-effect conclusions when significant differences are found between the two samples. In particular, confounding variables are also a possibility with observational studies. In describing potential confounding variables, make sure you explain the connections to both the explanatory variable and the response variable, and how the confounding variable provides an alternative explanation to the observed difference.

This led us to consider comparative experiments which utilize active imposition of the explanatory variable and random assignment to guard against confounding variables. In a properly designed experiment, which may include placebos, double-blindness, and other controls, there should not be any other substantial differences between the explanatory variable groups. Thus, if we do find a statistically significant difference in the response variable, we will feel comfortable attributing that difference to the imposed explanatory variable.

Then we returned to approximating p-values through simulation, but this time modeling the random assignment process under the null hypothesis. We “shuffled” the responses to the explanatory variable groups, and then calculated a statistic with each shuffle to assess the null distribution of that statistic under the null hypothesis. The hypergeometric probability distribution was seen again, this time as a mathematical model for this random assignment process (leading to Fisher’s Exact Test for calculating exact p-values). Again a normal probability model can also often be applied to approximate the p-value and to obtain confidence intervals. Keep in mind that the normal model should be applied only with suitably large sample sizes.

Lastly, you considered limitations to the difference in conditional probabilities as the parameter of interest and explored properties of relative risk and odds ratios as alternative ways to summarize the association between two categorical variables. In particular, with case-control studies, only the odds ratio provides a meaningful comparison. You applied your knowledge of simulation methods to obtain approximate p-values for these statistics, learning the usefulness of transformations in normalizing distributions. This allowed us to apply some normal-based methods for obtaining confidence intervals for the population relative risk and population odds ratio. Don’t forget to back-transform these intervals to return to the original scale.

**SUMMARY OF WHAT YOU HAVE LEARNED IN THIS CHAPTER**

- How to construct a two-way table
- Examining conditional proportions to explore group differences on a categorical response variable
- How to construct and interpret a segmented bar graph
- How to state hypotheses in terms of the difference in two population proportions
- When and how to use technology to apply two-sample  $z$  procedures to compare two sample proportions
- Interpretation of the  $z$ -confidence interval as the interval of plausible values for the *difference* in the population proportions or process probabilities
- Terminology of observational studies and experiments, including explanatory and response variables
- Being able to distinguish between observational studies and experiments
- Identifying and explaining potential confounding variables in observational studies
- How to carry out random assignment
- The benefits of random assignment
- Approximating p-values using a two-way table simulation
- Calculating exact p-values using the hypergeometric distribution
- Using the normal approximation to obtain p-values and confidence intervals for an underlying treatment effect (when and how)
- Factors that affect the statistical significance between the two groups (e.g., sample size, size of treatment effect)
- Distinctions between cohort, case-control, and cross-classified studies
- Distinctions (and equivalences) between relative risk and odds ratio (how to calculate, how to interpret)
- Confidence intervals for relative risk and odds ratios
- Limitations in the scope of conclusions that can be drawn from different study designs

**TECHNOLOGY SUMMARY**

- Creating segmented bar graphs (R, Minitab, Excel)
- Creating a null distribution of differences in sample proportions, relative risk, odds ratio
  - Simulating independent (binomial) random samples from a common population
  - Simulating random assignment using the Analyzing Two-way Tables applet and using the hypergeometric probability distribution
  - Simulating the random sampling associated with a case-control study
- Calculating hypergeometric probabilities
- Two-sample  $z$ -procedures

## Choice of Procedures for Comparing Two Proportions

Parameter	Difference in population proportions ( $\pi_1 - \pi_2$ )	Relative Risk ( $\pi_1/\pi_2$ )	Odds Ratio ( $\tau$ )
Study design	Two binary variables, but not case-control study	Two binary variables, but not case-control study	Two binary variables
Null Hypothesis	$H_0: \pi_1 - \pi_2 = 0$	$H_0: \pi_1/\pi_2 = 1$	$H_0: \tau = 1$
Simulation	<ul style="list-style-type: none"> <li>Independent random sampling from binomial processes</li> <li>Random assignment with hypergeometric distribution</li> </ul>		
Exact p-value	Fisher's Exact Test		
Can use z procedures if	At least 5 successes and 5 failures in each group		
Confidence interval	$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$\exp \left[ \ln(\hat{p}_1/\hat{p}_2) \pm z^* \sqrt{\frac{1}{A} - \frac{1}{A+C} + \frac{1}{B} - \frac{1}{B+D}} \right]$	$\exp \left[ \ln(\hat{\tau}) \pm z^* \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \right]$
R Commands	iscamtwopropztest <ul style="list-style-type: none"> <li><i>observed1</i> (either the number of successes or sample proportion for first group), <i>n1</i> (sample size for first group, <i>observed2</i> (count or proportion) and <i>n2</i> for the second group</li> <li><i>Optional: hypothesized difference</i> and <i>alternative</i> ("less", "greater", or "two.sided")</li> <li><i>Optional: conf.level</i></li> </ul>		fisher.test <ul style="list-style-type: none"> <li><i>matrix(c(A, C, B, D), nrow=2)</i></li> <li><i>alt=</i> ("less", "greater", or "two.sided")</li> <li><i>Optional:</i> <i>conf.int = TRUE</i>, <i>conf.level</i></li> </ul>
Minitab	Stat > Basic Statistics > 2 Proportions <ul style="list-style-type: none"> <li>Test: Use the pooled estimate of the proportions</li> <li>Interval: Estimate the proportions separately</li> </ul>		

**Note:** Fisher's Exact Test is usually considered a reasonable approximation to the p-value from independent random sampling.

## Quick Reference to ISCAM R Workspace Functions and Other R Commands

Procedure Desired	Function Name (Options)
Set up two-way table	<code>matrixname = (matrix, (c(A, C, B, D), nrow = 2, dimnames = list(c("row name", "row name"), c("column name", "column name"))))</code>
Segmented bar graph	Set up two-way table and then use barplot <code>barplot(prop.table(matrixname, margin=2), legend = T, ylab = "Proportion")</code>
Generate data from Binomial	<code>rbinom(number, sample size, probability)</code>
Create a column of 3 graphs	<code>par(mfrow=c(3,1))</code>
Two proportion z-test/interval	<code>iscamtwopropztest(x1, n1, x2, n2, hypothesized, alternative, conf.level) *x1 can be count or proportion</code>
Hypergeometric probability	<code>iscamhyperprob(k, total, succ, n, lower.tail = FALSE)</code>
Normal approximation to hypergeometric	<code>iscamhypernorm(x, N, M, n, TRUE/FALSE)</code>
Generate data from hypergeometric	<code>rhyper(number, sample size, probability)</code>
Create Boolean for values below some number $x$ (creating an empirical p-value)	<code>sum(data &lt;= x)/n</code>
Fisher's Exact Test, confidence interval for odds ratio	<code>fisher.test(matrix, alternative)</code>

## Quick Reference to Minitab Commands

Procedure Desired	Menu
Segmented bar graph	Graph > Bar Chart
Generate data from Binomial	Calc > Random Data > Binomial
Histogram	Graph > Histogram
Mean, standard deviation, sum of values	MTB> mean C1, MTB> std C1, MTB> sum C1
Summary statistics	Stat > Basic Statistics > Display Descriptive Statistics (or MTB> describe c1)
Overlay normal curve on histogram	Right click, Add > Distribution Fit
Normal Probability Plot	Graph > Probability Plot
Two proportion z-test/interval	Stat > Basic Statistics > 2 Proportions Use "not equal to" alternative for confidence interval
Hypergeometric probability	Graph > Probability Distribution Plot
Generate data from hypergeometric	Calc > Random Data > Hypergeometric
Create Boolean for values below some number $x$ (creating an empirical p-value)	Calc > Calculator, expression: <code>sum(c7 &lt;= x)/n</code> <code>MTB&gt; let c10 = (c7 &lt;= x)</code> <code>MTB&gt; let k1 = sum(10)/n</code> <code>MTB&gt; print k1</code>
Fisher's Exact Test	Stat > Basic Statistics > 2 Proportions

## CHAPTER 4: COMPARISONS WITH QUANTITATIVE VARIABLES

This chapter parallels the previous one in many ways. We will continue to consider studies where the goal is to compare a response variable between two groups. The difference here is that these studies will involve a *quantitative* response variable rather than a *categorical* one. The methods that we employ to analyze these data will therefore be different, but you will find that the basic concepts and principles that you learned in Chapters 1–3 still apply. These include the principle of starting with numerical and graphical summaries to explore the data, the concept of statistical significance in determining whether the difference in the distribution of the response variable between the two groups is larger than we would reasonably expect from randomness alone, and the importance of considering how the data were collected in determining the scope of conclusions that can be drawn from the study.

### Section 1: Comparing groups – Quantitative response

Investigation 4.1: Employment discrimination?

### Section 2: Comparing two population means

Investigation 4.2: NBA Salaries – Independent random samples,  $t$  procedures

Investigation 4.3: Left-handedness and life expectancy – Factors influencing significance

### Section 3: Comparing for two treatment means

Investigation 4.4: Lingering effects of sleep deprivation – Randomization tests

Investigation 4.5: Lingering effects of sleep deprivation (cont.) – Two-sample  $t$ -tests

Investigation 4.6: Ice cream serving sizes – Two-sample  $t$ -confidence interval

Investigation 4.7: Cloud seeding – Strategies for non-normal data

### Section 4: Matched Pairs Designs

Investigation 4.8: Chip melting times – Independent vs. paired design, technology

Investigation 4.9: Chip melting times (cont.) – Inference (simulation, paired  $t$ -test)

Investigation 4.10: Comparison shopping – Application

Investigation 4.11: Smoke alarms – McNemar's test (paired categorical data)

Example 4.1: Age Discrimination? – Randomization test

Example 4.2: Speed Limit Changes – Two-sample  $t$ -procedures

Example 4.3: Distracted Driving? (cont.) – Paired  $t$ -procedures

## SECTION 1: COMPARING GROUPS – QUANTITATIVE RESPONSE

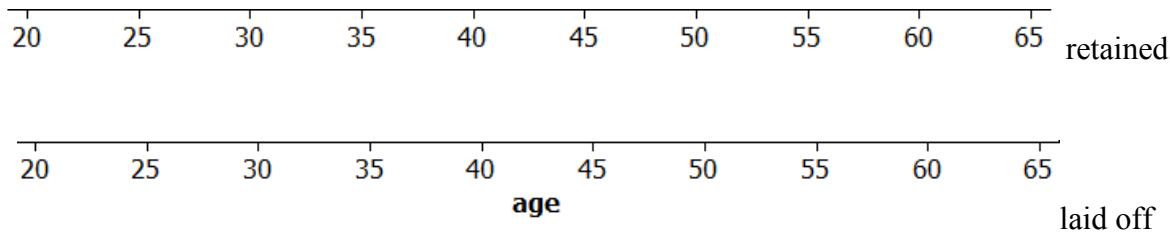
In this chapter, we will focus on comparing two groups on a quantitative response variable. Again, the reasoning is still the same as in Ch. 1 – 3, but we will see some small changes in the details. We start with a small investigation to review key ideas and then we parallel Chapter 3 in first considering random sampling and then random assignment, both for scope of conclusions but also for how we will design our simulation.

### Investigation 4.1: Employment Discrimination?

Robert Martin turned 55 in 1991. Earlier in that same year, the Westvaco Corporation, which makes paper products, decided to downsize. They ended up laying off roughly half of the 50 employees in the engineering department where Martin worked, including Martin. Later that year, Martin went to court, claiming that he had been fired because of his age. A major piece of evidence in Martin's case was based on a statistical analysis of the relationship between the ages of the workers and whether they lost their jobs.

Part of the data analysis presented at his trial concerned the ten hourly workers who were at risk of layoff in the second of five rounds of reductions. At the beginning of Round 2 of the layoffs, there were ten employees in this group. Their ages were 25, 33, 35, 38, 48, 53, 55, 55, 56, 64. Three were chosen for layoff: the two 55-year-olds (including Martin) and the 64-year old.

- (a) Create stacked dotplots of the ages of the employees who were retained and laid off.



Do these data seem to support the claim that the laid-off employees tended to be older than those who were not laid off?

- (b) One way to measure this support is to compare the difference in the average ages of the two groups. Calculate the mean age of the employees that were laid off and the mean age of the employees that were not laid off ( $\bar{x}_{\text{laid-off}} - \bar{x}_{\text{retained}}$ ).

- (c) Although we see some evidence of an age difference, on average, between these two groups, it's still possible that the employers just randomly decided which three employees to layoff. How would you expect the behavior of these dotplots to compare if the layoff decisions had been made completely at random?

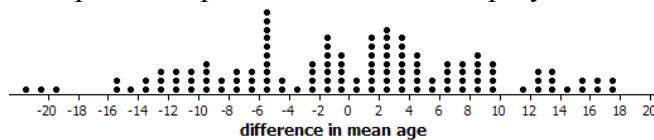
(d) Describe how you could generate one “could have been” observation under this assumption that the lay-off decisions are made completely at random.

(e) If we consider the outcomes where three of the ten employees are laid off, how many of these outcomes are “as extreme” as the observed result? How many are “more extreme”? [Hint: List the ages of the three employees that are laid off for these results.]

(f) How many different ways are there to form these two groups, three to be laid off and seven retained, if the decisions are made completely at random?

(g) How many of these random splits have three of the four oldest people in the laid-off group?

The dotplot below shows all possible values for the “difference in mean age” for the laid-off group minus the retained group for all possible splits into 3 laid-off employees and 7 retained.



(h) What is the approximate average of this distribution? Is this what you expected?

(i) Is this distribution symmetric? Why or why not?

**Discussion:** Although we might suspect the distribution to be centered at zero, the graph is not expected to be symmetric because the group sizes are unequal and we subtracted laid-off – retained. Also, the data set is small so there will be “granularity” between the possible values for the difference.

(j) Which of the outcomes in the dotplot correspond to the outcomes you listed in (g)? [Hint: Find the difference in group means when the following ages are assigned to the laid-off group.]

55, 55, 56:

55, 55, 64:

55, 56, 64:

53, 56, 64:

(k) Use your result from (f) and (g) to calculate an exact p-value for the null hypothesis that the lay-off decisions were made completely at random. How can you interpret this p-value? [Hint: How was randomness used in the study?]

(l) What do you conclude from the p-value you calculated in (k)? [Hint: Think about the significance of the result and the type of study.]

(m) Remind the lawyers what limitations there are to the conclusions that can be made in this study.

### Study Conclusions

In this analysis, you assumed a null model – the lay-offs are made completely at random – and you considered all 120 possible assignments of 7 people to be retained and 3 people to be laid off. (You could have flipped a coin for each employee, but that would lay off 50% in the long-run rather than fixing 3 to be laid off.) Of these 120 possible assignments, we want to know how many are “at least as extreme” as the observed result. If we focus on the outcomes where 3 of the 4 oldest people end up in the laid-off group, only the 55, 56, 64 combination has a difference in means (17.33 years) larger than what we observed (16.86, which can happen two ways). The 55, 55, 56 combination has a smaller difference in means (13.04) and the 53, 56, 64 combination also does not quite have as large a difference in means (16.38). So the exact p-value = 3/120 = 0.025, which tells us that if these decisions had been made completely at random, there is a 2.5% chance we would see a combination of ages at least as extreme as the actual outcome as measured by the difference in group means. This is moderate evidence that the decision making process was not at random. If you do find this evidence compelling, you must keep in mind the observational nature of the data; we cannot draw any cause-and-effect conclusions about the reasons for the non-random lay-off decisions.

The reasoning process of statistical significance remains the same – how often would random chance give me a result at least as extreme? We will still need to consider how we want to measure the group differences (e.g., choice of statistic) and how we can approximate the p-value (e.g., by modelling the *hypothetical* randomness assumed by the null hypothesis), especially when a listing of all possible outcomes is not feasible. Large sample sizes will also again allow us to use a convenient confidence interval formula. Once you have a p-value or confidence interval, you interpret them in the same way as in previous chapters, while also considering the limitations of the study design.

### Practice Problem 4.1

- (a) Suppose the following individuals were laid off: 25, 35, 38. What would be the difference in the mean age of those who were laid off and those who were retained?
- (b) What would be the exact p-value for this difference in means from part (a) if the alternative hypothesis was that there was a bias against the younger employees?

## SECTION 2: COMPARING TWO POPULATION MEANS

In this section, we want to focus on quantitative data arising from two independent random samples. We will first begin with a hypothetical situation to explore how our statistic should behave under certain assumptions about the populations. We will then discuss a theoretical result justifying use of a *t*-distribution for the reference distribution in finding p-values and confidence intervals.

### Investigation 4.2: NBA Salaries

The file [NBASalaries2014.txt](#) contains season salaries (in millions of dollars) for all NBA basketball players at the start of the 2014-15 season (posted Oct 16, 2014; downloaded from [BallnRoll](#), July 2015). Players without a team affiliation or without a listed salary were not included in the data set.

(a) Do you expect salary data to be symmetric or skewed, and if skewed, in which direction? Explain.

(b) Which do you suspect will be larger, the mean or median salary? Explain.

### Descriptive Statistics

When using graphs to compare groups it is especially important that *comparative graphs* be drawn on the same scale.

#### Technology Detour – Comparative Graphs

##### In Descriptive Statistics applet

Copy the salary and conference data to the clipboard. In the applet, check the **Stacked** box and confirm the ordering of the category and quantitative variables. Paste the data into the Sample data window and press **Use Data**. You have radio button options for Histograms and Boxplots.

##### In R

```
> iscamdotplot(time, year)           ← Input quantitative then categorical variable
> histogram(~salary | conference,  ← Remember to load lattice package
+ layout = c(1,2), nint=20)         ← Creates vertical layout, number of bins
> iscamboxplot(salary, conference) OR
> boxplot(salary ~ conference, horiz=TRUE, ylab="salary (mil $)")
```

And we can easily calculate separate summary statistics for the two conferences:

### Technology Detour – Numerical Summaries

#### In R

To get separate descriptive statistics for each group:

```
> iscamsummary(salary, conference)
```

Note: This adds the second (explanatory) variable. You still have the option of entering “digits =” to specify the number of significant digits you want displayed.

(c) Create a comparative graph and calculate the difference in the population mean salaries between the Eastern Conference and the Western Conference. (Notice that because the dataset lists salaries for all players in the NBA, you are working with a population and not a sample.) What symbols should you use to reference to this difference?

(d) Also determine the population sizes and the population standard deviations and record these below, using appropriate symbols.

(e) Suppose we weren’t able to perform a *census* and only had resources to find the salaries for 40 players. How would you select the 40 players?

(f) Use technology to select a random sample of 20 players from each league:

### Technology Detour – Selecting Independent Random Samples

#### In R take a random sample from each division

```
> westsample = sample(salary[which(conference == "western")], 20)
> eastsample = sample(salary[which(conference == "eastern")], 20)
> salary.sample = data.frame(westsample, eastsample)
> names(salary.sample) = c("western", "eastern")
```

(g) Create comparative graphs and descriptive statistics for these sample data. [Note: Your data is currently “unstacked” which requires a different approach than “stacked data” in most software. See Technology Detour on next page.]

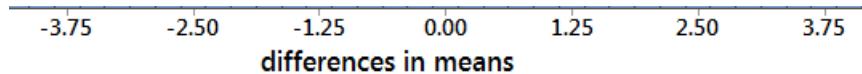
### Technology Detour – Unstacked data

In R have to first stack the data

```
> StackedData = stack(salary.sample[, c("western", "eastern")])
> names(StackedData) = c("salary", "division")
> iscamdotplot(StackedData$salary, StackedData$division)
> iscamsummary(StackedData$salary, StackedData$division)
```

Record the sample means and sample standard deviations for your samples. How do the sample results compare to the population results? What about the shapes of the samples?

(h) Calculate your difference in means and pool with the rest of your class.



(i) How would you describe the shape of this distribution? Is this what you expected?

(j) What is the approximate center of this distribution? Is this what you expected?

(k) But of course, what you really want to know about is the variability. Is there more or less variability in this distribution compared to the variability in each population (part (d))?

(l) We know there should be less variability because we are talking about sample means rather than individual observations. But how much variability do we expect to find in the distribution of sample means taken from the Western conference? What about the Eastern conference? Is the variability in the difference in sample means larger or smaller than in the individual distributions?

(m) To confirm these observations, we need to take a lot more samples. Outline the simulation steps we need to use.

(n) Use a premade script or macro to carry out the simulation:

- In R: Run the [NBASalarySamples.R](#) script to generate 1000 repetitions.
- 

Calculate the *difference* in sample means for each repetition. Examine and describe the distribution of the *differences* of sample means. [Hint: Reconsider questions (i), (j), and (k).]

**Key Result:** A theorem similar to the Central Limit Theorem indicates that if we have two infinite, normally distributed populations, with means  $\mu_1$  and  $\mu_2$  and standard deviations  $\sigma_1$  and  $\sigma_2$ , and we take independent random samples from each population, the sampling distribution of  $\bar{X}_1 - \bar{X}_2$  will follow a normal distribution with mean  $\mu_1 - \mu_2$  and standard deviation  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ .

This is a nice theoretical result, but we seldom meet these conditions exactly.

(o) Identify two characteristics of our populations that do not agree with this statement.

A third, but perhaps the biggest, issue is that the population standard deviations are unknown.

(p) Suggest a formula for estimating the standard deviation of the difference in sample means.

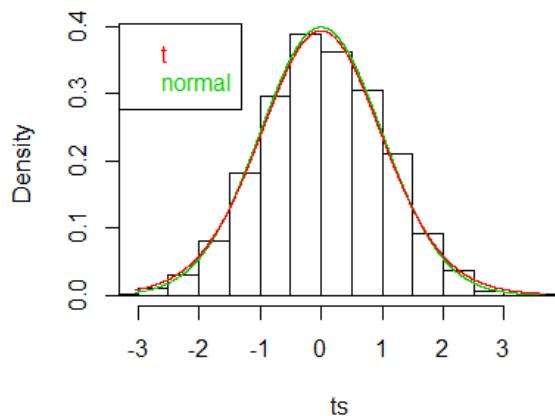
So then what we really need to know about is the distribution of  $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ .

(q) Create this statistic for each repetition of your earlier simulation results and then examine the distribution of this statistic. Is it approximately normal? Where is it centered?

**Discussion:** The sample sizes of 20 for each group appear to be large enough for a *t*-distribution (even a normal distribution) to reasonably approximate the distribution of this statistic. If the population distributions had themselves been normally distributed (and large), then a *t*-distribution will be a good approximation of the statistic for any sample sizes. The more skewed the population shapes, the larger the sample sizes need to be for you to safely use the *t*-distribution. Notice even with these very skewed populations, because their shapes and sample sizes were similar, we didn't need very large sample sizes at all for the standardized statistic to follow a *t* distribution. One remaining question is what to use for the degrees of freedom for the *t*-distribution. In the one-sample case we used  $n - 1$ . For two samples, it's actually more complicated. There is a "Welch-Satterwaite" approximation that involves the sample sizes and the unknown population standard deviations, though it doesn't always result in integer values. We will leave this calculation up to technology. You only need to realize that the degrees of freedom are related to the sample sizes and as the sample sizes becomes large, the *t*-distribution loses more and more of the heaviness in the tails and approaches the (standard) normal distribution. If you need to determine a p-value or critical value "by hand," we suggest using the following simple approximation for the degrees of freedom:  $\min(n_1 - 1, n_2 - 1)$ , the smaller of the two sample sizes minus one.

In this case, we could use 19 as the approximate degrees of freedom. Notice that this sample size is large enough that the *t*-distribution looks a lot like the normal distribution as well, which should agree with your observation in (q). But we will continue to use the *t*-distribution rather than the normal distribution to still account for additional variation from our estimates of the population standard deviations.

Especially when the population shapes are similar and the sample sizes are similar, the *t*-distribution provides a pretty good approximation to the sampling distribution of  $\bar{X}_1 - \bar{X}_2$ , even for non-normal populations.



## Summary of two-sample *t* procedures

**Parameter:**  $\mu_1 - \mu_2$  = the difference in the population means

**To test  $H_0: \mu_1 - \mu_2 = \delta_0$**

$$\text{Test statistic: } t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

***t*-Confidence interval for  $\mu_1 - \mu_2$ :**

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Approximate degrees of freedom:  
Compare this to a *t*-distribution with degrees of freedom equal to the smaller of the two sample sizes minus one:  
 $\min(n_1, n_2) - 1$ .

**Technical conditions:** These procedures are considered valid if the sample distributions are reasonably symmetric or the sample sizes are both at least 20.

(r) Use your original samples from (g) and carry out a two-sample *t*-test (by hand or see the Technology Detour below for summary statistics) of whether the population means are equal.

(s) Also calculate and interpret a 95% confidence interval. [Hint: When interpreting a confidence interval for the *difference*, you should always clarify the direction of subtraction and which population parameter is larger.]

(t) Do you have strong evidence of a difference in the population means? Is this what you would expect? Explain.

### Study Conclusions

In this study we conducted a census of the salaries of all NBA players prior to the start of the 2014-15 season. We find that the mean salaries are essentially the same,  $\mu_{\text{eastern}} = \mu_{\text{western}} = \$4.261$  million. The standard deviations are quite large compared to these mean values ( $\sigma_{\text{eastern}} = \$4.66$  million and  $\sigma_{\text{western}} = 4.73$  million) and the population distributions are strongly skewed to the right. If we had only sampled 20 players from each league, independently, then we would expect the difference in sample

means to also be close to zero, with a standard deviation of  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{4.66^2}{20} + \frac{4.73^2}{20}} \approx 1.48$ .

Of course we wouldn't usually know that, because we are unlikely to know the population standard deviations. However, we could estimate this quantity using the sample standard deviations. In doing so, we will again use the *t* distribution as our reference distribution for our statistic with quantitative data. In this case, because the null hypothesis is nearly true, we would expect to fail to reject the null hypothesis from your individual samples, and we expect about 95% of your confidence intervals to include zero. Suppose we had found a confidence interval of  $(-5.13, -0.63)$ . Then you would interpret this interval by saying you were 95% confident that the mean salary for all eastern conference players is \$0.63 to \$5.13 million dollars smaller than the mean salary for all western conference players.

### Practice Problem 4.2

Repeat the above simulation analysis using sample sizes of 10 for each conference. Write a summary of your observations.

## Using Technology to Carry out the Two-Sample $t$ -Tests

In using technology, you need to consider how the data are available: Do you have the “raw data” (meaning all the individual data values) or just the summary statistics (sample sizes, means, SDs)? Most software packages will assume raw data is “stacked” – each column represents a different variable (all response variable values in one column, the explanatory variable in a second column). Also notice that being able to run the  $t$ -test with only the summary data is one advantage the  $t$ -test has over the simulation. But also keep in mind that the  $t$ -output is valid only when the population distributions are normal or the sample sizes are large. When reporting your results, make sure you specify which technology you used, as different software use different degrees of freedom. You will also need to watch for the direction of subtraction that the software uses automatically (often alphabetical by group name).

There is also a version of the two-sample  $t$ -test that makes an additional assumption – the population standard deviations are equal. The standardized statistic has the following form:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}} \text{ where } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \text{ a pooled estimate of the common SD.}$$

This “[pooled  \$t\$ -test](#)” has some advantages (namely higher power), when the population standard deviations are equal. However, this additional condition can be difficult to assess from your sample data. The benefits do not appear to outweigh the risks of applying this assumption when you shouldn’t. So in this text we will focus on “unpooled  $t$ -tests” only.

### Technology Detour – Two-Sample $t$ -Test (unpooled, summary data)

**In R** The `iscamtwosampl` function takes the following inputs:

- Group 1 *mean, standard deviation, sample size* (in that order)
- Group 2 *mean, standard deviation, sample size* (in that order)
- *Optional: hypothesized difference* (default is zero)
- *Optional: alternative* (“less”, “greater”, or “two.sided”)
- *Optional: conf.level*

For example: `iscamtwosampl(x1=75, sd1=15, n1=888, x2=66, sd2=15, n2=99, alt="greater")`

#### Theory-Based Inference applet

- Use the pull-down menu to select “Two means.”
- Specify the summary statistics. Press **Calculate**. (You will see a visual representation of the sample means and standard deviations.)
- Check the box for Test of Significance. Specify the hypothesized difference and the direction of the alternative hypothesis.
- Press **Calculate**.

Scenario: Two means

Paste Data

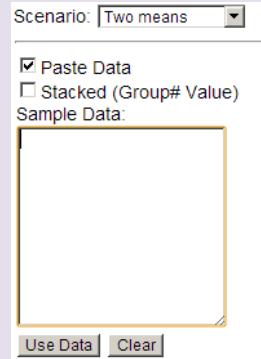
Group 1	Group 2
n: 888	n: 99
mean, $\bar{x}$ : 75	mean, $\bar{x}$ : 66
sample sd, s: 15	sample sd, s: 15

Calculate

## Technology Detour - Two-Sample $t$ -Test (unpooled, raw data)

### Theory-Based Inference applet

- Use the pull-down menu to select “Two means.”
- Copy the data to the clipboard, check the **Paste data** box, select the **Stacked** box if that’s the data format (explanatory variable in first column), press **Clear**, click in the data window and paste the data. Press **Use data**.
- Check the **Test of Significance** box.
- Specify the hypothesized difference and the direction of the alternative hypothesis.
- Press **Calculate**.



In R (note, if you have attached the data, you don't need the “with”)

```
> with(NBASalaries2014,  
+       t.test(salary~conference, alt="two.sided", var.equal=FALSE))
```

### Investigation 4.3: Left-Handedness and Life Expectancy

Psychologist Stanley Coren has conducted several studies investigating the life expectancy of left-handers compared to right-handers, believing that the stress of being left-handed in a right-handed world leads to earlier deaths among the left-handers. In one study Coren and Halpern (1991) sent surveys to thousands of next-of-kin of recently deceased southern Californians and asked whether the person had been right-handed or left-handed. They were very careful in how they collected their data. First, they consulted a bereavement counselor who suggested that they not contact anyone unless at least 9 months had passed since the death. The counselor also suggested that they make the contact as gentle as possible and not follow up or press people for responses. The researchers also decided that they would not contact next of kin if the death had been a result of murder or suicide or if the deceased was a child age 6 or younger. They received 987 replies and found that the average age of right-handed people at death was 75 years and for left-handed people it was 66 years.

(a) Is this an observational study or an experiment?

(b) Did the researchers take a random sample of left-handers and an independent random sample of right-handers? Is it reasonable to consider these samples independent?

Even though the researchers did not take a random sample of left-handers and a separate random sample of right-handers, we are still willing to consider these samples as independent because the results for one group should have no effect on the results for the other group. In a situation with two independent random samples, we can apply the two-sample  $t$ -test to make inferences about the difference in the population means.

(c) The summary of the study reported the sample means; what additional information do we need to assess the statistical significance of their difference? Why is this information important?

(d) Let  $\mu_L$  represent the mean lifetime for the population of left-handers and let  $\mu_R$  represent the mean lifetime for the population of right-handers. State the null and alternative hypotheses for Coren and Halpern's study (based on the research conjecture).

The table below lists some guesses for the sample size breakdown (based on estimates of the proportion of the population that are left-handed) and the sample standard deviations.

Scenario		Sample sizes	Sample means	Sample SDs	t-statistic	p-value	Significant at 1% level?
1	left	99 (10% of 987)	66	15			
	right	888	75	15			
2	left	50 (5% of 987)	66	15			
	right	937	75	15			
3	left	50 (5% of 987)	66	25			
	right	937	75	25			
4	left	10 (1% of 987)	66	25			
	right	977	75	25			
5	left	99 (10% of 987)	66	50			
	right	888	75	50			

(e) For the following pairs of scenarios from the table above, which do you believe will result in the lower p-value, that is, stronger evidence against the null hypothesis? Explain briefly.

Scenario 2 vs. 3

Scenario 3 vs. 4

(f) Use technology (see preceding Technology Detour) to carry out the two-sample *t*-tests for these hypotheses, using the sample sizes, sample means, and sample standard deviations given in the table above. For each of these five scenarios, report the resulting test statistic, p-value, and whether or not the difference is statistically significant with a significance level of  $\alpha = 0.01$  in the table above.

(g) Summarize what your analysis reveals about the effects of the sample size breakdown and the sample standard deviations on the values of the *t*-statistic and p-value.

(h) Considering the five scenarios of sample sizes and standard deviations used in the table, which scenario do you think is most realistic for this study of left- and right-handers' lifetimes? Explain, based on the context for these data.

(i) Based on your analyses and how these data were collected, are you convinced that being left-handed causes individuals to die sooner on average than being right-handed? How would phrase your conclusion?

(j) Part of the motivation for this research was an earlier study by Porac and Coren (1981) that surveyed 5147 men and women of all ages in North America. They found that 15% of ten-year-olds were left-handed, compared to only 5% of fifty-year-olds and less than 1% of eighty-year-olds. At the age of 85, right-handers outnumbered left-handers by a margin of 200 to 1. Suggest another explanation for these puzzling findings that actually provides a counter-argument to the conclusion from the 1991 Coren and Halpern study that left-handers tend to live shorter lives than right-handers.

### Study Conclusions

The difference in the sample mean lifetimes does appear to be statistically significant for all reasonable choices of the sample sizes and the sample standard deviations, so we have strong evidence that left-handers do tend to have shorter life spans than right-handers. However, there are numerous cautions to heed when drawing conclusions from such a study. This was a retrospective study with voluntary response, and in fact the researchers reported that they tended to hear from the left-handed relatives more often than right-handers, and they only heard from fewer than half of the families contacted. There is also no information given about the proportion of left-handers in the two southern California counties studied or the average ages of their residents now. In particular, one explanation for the lower percentage of left-handers among the elderly is not that they have died younger but that it used to be quite common practice to strongly encourage left-handed children to switch to being right-handed. Nowadays, that is less common (in fact many athletes love having this advantage!), and so there is a higher percentage of left-handers among younger age groups. This helps to explain why the left-handers who had died would tend to be younger. In other words, maybe the average age difference between *living* left-handers and right-handers is also nine years.

These studies have actually become hot topics for debate as some other studies have not been able to replicate Coren and Halpern's results. Other prospective longitudinal studies in the United States (Marks and Williams, 1991; Wolf, D'Agostino, and Cobb, 1991) have not found a significant difference in age at death. Still others have found connections between handedness and accident rates, lower birth rates, cancer, alcohol misuse, smoking, and schizophrenia. Alas, we don't see any randomized, comparative experiments being conducted to answer these questions in the near future!

**Discussion:** You should have found that larger variability in each sample (so a larger SE) produces a larger p-value and therefore less convincing evidence that the population means differ. Or to put this in a more positive light: Reducing variability within groups makes it easier to distinguish between the groups. You should also have found that a bigger discrepancy in sample sizes between the two groups produces a larger p-value and therefore less convincing evidence that the population means differ.

(k) If all else stays the same but the difference in the means between the groups is larger, will the p-value be larger or smaller?

(l) If all else stays the same but the sample sizes within the groups are larger, will the p-value be larger or smaller?

### Practice Problem 4.3

Recall the study on children's television viewing habits for two schools in San Jose, CA from Practice Problem 2.5B. The researchers wanted to study whether a new classroom curriculum could reduce children's television viewing habits, which might in turn help to prevent obesity. One of the schools, chosen at random, incorporated an 18-lesson, 6-month classroom curriculum designed to reduce watching television and playing video games, whereas the other school made no changes to its curriculum. Both before the curriculum intervention, all children were asked to report how many hours per week they spent on these activities.

The following summary statistics pertain to the reports of television watching, in hours per week, *prior* to the intervention:

Baseline	Sample size	Sample mean	Sample SD
<b>Control group</b>	103	15.46	15.02
<b>Intervention group</b>	95	15.35	13.17

Approximate two-sided p-value: 0.956

Suppose that the summary statistics at baseline had instead been the following:

	Sample size	Sample mean	Sample SD
<b>Control group</b>	103	14.46	15.02
<b>Intervention group</b>	95	8.80	13.17

(a) Without performing any calculations, how should the p-value for these data compare (larger or smaller) to the p-value from the actual baseline results? Explain.

Suppose that the summary statistics at the beginning of the study had instead been the following:

	Sample size	Sample mean	Sample SD
<b>Control group</b>	103	15.46	13.82
<b>Intervention group</b>	95	15.35	10.41

(b) Without performing any calculations, how should the p-value for these data compare (larger or smaller) to the actual baseline results? Explain.

(c) Suppose the researchers decide to look at a subset of children in this study that belong to the same social-economic class (with the expectation that their television watching habits will be more similar to each other). Discuss one advantage and one disadvantage to this approach in terms of detecting a difference between the control group and the intervention group at the conclusion of the study. [Hints: Power? Generalizability?]

### SECTION 3: COMPARING TWO TREATMENT MEANS

The previous section focused on quantitative data arising from two independent random samples. This section will focus on quantitative data arising from randomized experiments. We will again consider simulation-based, exact, and theory-based p-values, as well as what assumptions we need to make for confidence intervals.

#### Investigation 4.4: Lingering Effects of Sleep Deprivation

Researchers have established that sleep deprivation has a harmful effect on visual learning (the subject does not consolidate information to improve on the task). In a recent study Stickgold, James, and Hobson (2000) investigated whether subjects could “make up” for sleep deprivation by getting a full night’s sleep in subsequent nights. This study involved randomly assigning 21 subjects (volunteers between the ages of 18 and 25) to one of two groups: one group was deprived of sleep on the night following training with a visual discrimination task, and the other group was permitted unrestricted sleep on that first night. Both groups were allowed unrestricted sleep on the following two nights, and then were re-tested on the third day. Subjects’ performance on the test was recorded as the minimum time (in milliseconds) between stimuli appearing on a computer screen for which they could accurately report what they had seen on the screen. Previous studies had shown that subjects deprived of sleep performed significantly worse the following day, but it was not clear how long these negative effects would last. The data presented here are the *improvements* in reaction times (in milliseconds), so a negative value indicates a decrease in performance.

Sleep deprivation group ( $n = 11$ ):  $-10.7, 4.5, 2.2, 21.3, -14.7, -10.7, 9.6, 2.4, 21.8, 7.2, 10.0$

Unrestricted sleep group ( $n = 10$ ):  $25.2, 14.5, -7.0, 12.6, 34.5, 45.6, 11.6, 18.6, 12.1, 30.5$

#### Study Design

(a) Is this an experiment or an observational study? Explain.

(b) Identify the explanatory (EV) and response (RV) variables. Also classify each as being categorical or quantitative.

EV:

type:

RV:

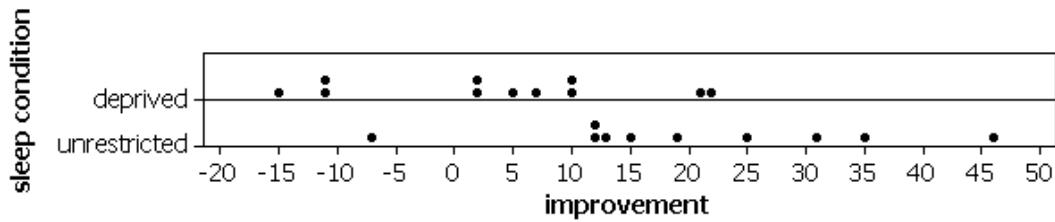
type:

(c) Write out the null and alternative hypotheses, in words, for this research study.

(d) Is your alternative hypothesis one-sided or two-sided? What does this imply about the types of response values you would expect to see in each treatment group? [Hint: Positive or negative?]

### Descriptive Statistics

The following dotplots reveal the distributions of improvements between the two groups:



(e) Do these data provide preliminary evidence that the harmful effects of sleep deprivation linger three days later? Explain your reasoning. [Hints: Are all the improvement scores in one group higher than all the improvement scores in the other group? Is there a tendency for higher improvement scores in one group three days later?]

(f) There are several ways we could choose to measure the tendency observed in (e). A common choice of statistic is of course the difference in group means. Calculate this statistic by subtracting the “deprived” group’s value from the “unrestricted” group’s value.

Observed difference in sample means:  $\bar{x}_{\text{unrestricted}} - \bar{x}_{\text{deprived}} =$

(g) Is it *possible* that the differences seen here could have occurred just by chance variation, due to the random assignment of subjects to groups, even if there were really no effect of the sleep condition on improvement?

(h) As usual, we want to know how *plausible* it is for us to obtain a difference in sample means at least as extreme as 15.92 by chance alone. What is the source of “chance” in this study?

(i) Outline the steps of a simulation analysis that would explore the “could have been” outcomes under the null hypothesis by modelling the randomness described in (h).

### Statistical Inference

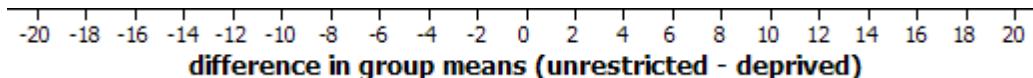
As with the “dolphin therapy” experiment from Chapter 3, we again need to judge the strength of evidence that the experimental data provide in support of the researchers’ conjecture that sleep deprivation has a harmful effect on learning. We are now working with a quantitative response variable rather than a categorical one, but we will use the same basic logic of *statistical significance*: We will ask whether the observed experimental results are very unlikely to have occurred by chance variation if the explanatory variable has no effect, that is, if the values in the two groups are interchangeable. To simulate this *randomization test*: We will take the observed response variable outcomes, randomly assign these numerical values between the two groups, compute the statistic of interest (e.g., the difference in group means), repeat this process a large number of times, and see how often the random assignment process alone produces a difference in group means as extreme as in the actual research study. This is similar to what you considered in Investigation 4.1. However, this time it is not really feasible to list out all possible  $C(21,10)$  random assignments and consider the value of the statistic for each one, so we will simulate a large number of random assignments instead. Note, we need to know the value of  $(\bar{x}_1 - \bar{x}_2)$ , not just the “number of successes” as with the dolphin therapy study.

- (j) Take a set of 21 index cards, and write each of the improvement values on a card. Then shuffle the cards and deal out 10 of them to represent the subjects randomly assigned to the “unrestricted sleep” group and 11 to represent the “sleep deprived group.” Calculate the mean of the improvements in each group. Then calculate the difference in group means, subtracting in the same order as before.

Simulated difference in means ( $\bar{x}_{\text{unrestricted}} - \bar{x}_{\text{deprived}}$ ):

- (k) Is your simulated result for the difference in means as extreme as the actual experimental result? Explain. Why is looking at this one simulated difference not enough to assess statistical significance?

- (l) Combine your results with your classmates to produce a dotplot of the *difference in group means*.



Simulating more repetitions would provide a better understanding of how significant (i.e., unlikely to have happened by random assignment alone) the observed experimental results are. In other words, more repetitions will enable us to approximate the p-value more accurately. We will again turn to technology to perform the simulation more quickly and efficiently.

(m) Open the [Comparing Groups \(Quantitative\)](#) applet.

You will see dotplots of the research results.

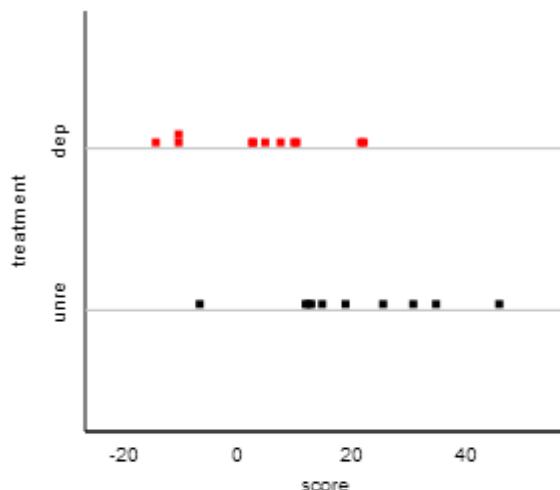
Verify the calculation of the observed difference in group means. (This applet subtracts the top row from the bottom row.)

- Check the **Show Shuffle Options** box.
- Select the **Plot** radio button.
- Press the **Shuffle Responses** button.

The applet combines all the scores into one pile, reassigned the group labels (red and black coloring) at random, and then redistributes the observations to the two treatment groups, just like you did with the index cards.

What did you find for the difference in group means this time?

Is this more extreme than the observed difference in group means in the research study?



#### Summary Statistics:

	n	Mean	SD
dep	11	3.90	12.17
unre	10	19.82	14.73
pooled	21	11.48	13.44

(n) Press **Shuffle Responses** four more times. With each new “could have been” result (each new random assignment), the applet calculates the difference in group means and adds a dot to the dotplot to the right. Identify what the observational units and variable are for the dotplot on the far right.

Obs. units:

Variable:

(o) As we conduct more and more repetitions, what do you expect the mean of this null distribution of difference in group means to be close to? Explain.

(p) Change the **Number of repetitions** from 1 to 995 (to produce a total of 1000 repetitions). Press **Shuffle Responses**. Describe the shape, mean, and standard deviation of the resulting dotplot. Explain what is represented by this standard deviation.

Shape:

Mean:

Standard deviation:

Interpretation:

(q) Now enter the observed difference in group means (question (f)) from the research study into the **Count Samples** box, use the *Greater Than* option (to match our alternative hypothesis), and press the **Count** button. What does the applet report for the empirical p-value?

(r) Provide an interpretation of this p-value in context (make sure you address the statistic, the source of the randomness in the study, and what you mean by “more extreme”).

### Conclusions

(s) *Significance*: What conclusion would you draw from this simulation analysis regarding the question of whether the learning improvements in the sleep deprived group are *significantly* lower (on average) than those in the unrestricted sleep group? Also explain the reasoning process by which your conclusion follows from the simulation results.

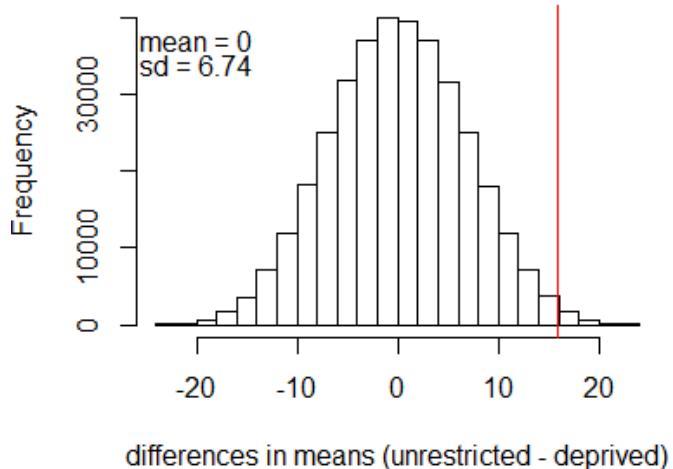
(t) *Causation*: Does the design of this study allow you to conclude that the reduction in improvement scores is caused by the sleep deprivation? Explain.

(u) *Generalizability*: To what “population” is it reasonable to generalize these results?

### Exact p-value?

Although not as convenient as before, we can determine the exact p-value for this randomization test by considering all the different possible random assignments of these 21 subjects into groups of 11 and 10, determining the difference in means (or medians) for each, and then counting how many are at least as extreme as the observed difference.

The following histogram (produced using the `combn` function in R, from the `combinat` package, assuming `imprvs` contains the response variable values) shows the distribution of all 352,716 possible differences in group means.



```

indices = 1:21
allcombs = combn(21, 10)

diffs = 1:ncol(allcombs)

for (i in 1:ncol(allcombs)){
  group1=imprvs[,i]
  group2=imprvs[setdiff(indices,
    allcombs[,i])]
  diffs[i]=mean(group1)-mean(group2)
}

hist(diffs)
abline(v=15.92, col=2)

```

approx. run time: 1 min

(v) Do your simulation results reasonably approximate this null distribution?

(w) It turns out that 2456 of the 352,716 different random assignments produce a difference in group means of 15.92 milliseconds or larger. Use this information to determine the exact p-value of this randomization test. Is the approximate p-value from your simulation close?

**Discussion:** The exact randomization distribution consists of every possible random assignment and calculates the statistic of interest (e.g., difference in means) for each one. Then we simply count how many of the configurations result in a value of the statistic at least as extreme (as defined by the alternative hypothesis) as the actual observed result. As you might expect, it can be extremely tedious, even with computers, to list out all of these possible random assignments. And these group sizes are relatively small! One shortcut is to only count how many assignments give results more extreme than the one observed, but as you will see we can often appeal to a mathematical model as well.

Note: You can carry out these simulations in R or Minitab as well, see the Technology Detour on the next page.

## Technology Detour – Simulating a Randomization Test for a Quantitative Response

### In [Comparing Groups \(Quantitative\)](#) applet

Copy and paste data (match the ordering of the explanatory and response variables) and press Use Data. Check the Show Shuffle Options box, specify the number of shuffles and press Shuffle Responses. Enter the observed statistic in the Count Samples box, choose a direction (less than, greater than, or beyond), and press Count.

### In R (with Sleep Deprivation file loaded and attached)

In the R Console, set the number of repetitions and initialize the vector that will store the differences.

```
> I = 10000           ← use upper case
> diff = 0           ← initializes the vector
```

Then create a loop that will mix up the order of the response variable values and calculate a difference in means with each new random assignment:

```
> for (i in 1:I){
+   rerandom = sample(improvement)
+   boxplot(rerandom~condition,
+   horizontal=TRUE)
+   diff[i]=mean(rerandom[1:10]) -
+     mean(rerandom[11:21])
> }
> Then you can examine a histogram of diff and compute the
> empirical p-value as before.
> hist(diff); abline(v =15.92, col=2)           ← adds vertical line
> sum(diff >= 15.92)/I                         ← p-value
```

← the loop for I iterations  
 ← mixes up response values  
 match up the explanatory  
 variable group sizes

### Study Conclusions

These data come from a randomized, comparative experiment. The dotplots and descriptive statistics reveal that, even three days later, the sleep-deprived subjects tended (on average) to have lower improvements than those permitted unrestricted sleep. To investigate whether this difference is larger than could be expected from random assignment alone (assuming no real difference between the two treatment conditions, our null hypothesis), you simulated a **randomization test** by assigning the 21 measurements (improvement scores) to the two groups at random. You should have found that random assignment alone rarely produced differences in group means as extreme as in the actual study (the “exact” p-value is less than 0.01). Thus, we have fairly strong evidence that the average learning improvement is genuinely lower for the sleep deprived subjects. Moreover, because this was a randomized comparative experiment and not an observational study, we can draw a causal conclusion that the sleep deprivation was the cause of the lower learning improvements. However, the subjects were college-aged volunteers, so we may not want to generalize these results to a much different population.

### Practice Problem 4.4A

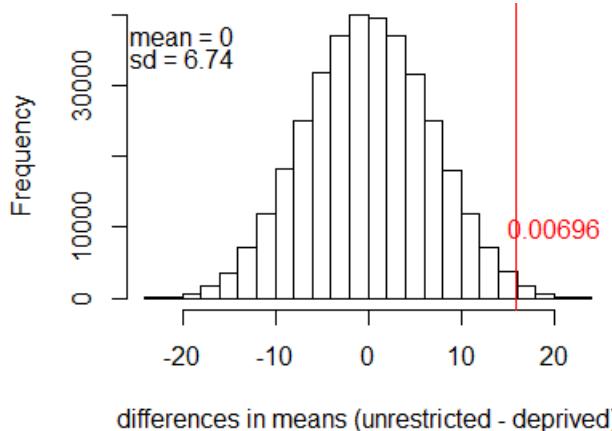
Explain how the simulation analysis conducted in this section differs from that of the previous section. For example: What is the random process being simulated? What assumptions are underlying the simulation?

### Practice Problem 4.4B

- Produce numerical and graphical summaries for the data in [FakeSleepDeprivation.txt](#), representing a new set of 21 responses for the sleep deprivation study. Comment on how the shapes, centers, and variability for the two distributions compare between these data and the original data.
- Use technology (see Technology Detour on previous page) to carry out a randomization test to compare the improvements for the sleep deprived and unrestricted sleep groups using the hypothetical data. Indicate how you approximated the p-value.
- How does the p-value for the hypothetical data compare to the p-value for the original data? Explain why this makes sense based on what you learned about how the data sets compared in (a).

### Investigation 4.5: Lingering Effects of Sleep Deprivation (cont.)

Reconsider the exact randomization distribution of the differences in sample means for the Sleep Deprivation study. Previously you determined simulation-based p-values and the exact p-value. Now we will explore modeling the randomization distribution of the difference in sample means with a probability distribution.



- (a) Does this distribution appear to be well modeled by a normal distribution?
- (b) Suppose you want to use the normal model to approximate the p-value for obtaining a difference in means of 15.92 or larger under the null hypothesis or to compute a confidence interval. What other information do you need to know?

To use the normal distribution, we need a measure of the variability in the randomization distribution. In Investigation 4.2, we found that when we have independent random samples from infinite populations we can use

$$SD(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where  $\sigma$  represents the “population” standard deviation. Because those  $\sigma$  values will almost surely be unknown, we can estimate them using the sample standard deviations, producing the “standard error” of the difference in sample means:

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- (c) For the sleep deprivation data, the sample standard deviations are:

$$s_{\text{unrestricted}} = 14.72 \text{ ms} \text{ and } s_{\text{deprived}} = 12.17 \text{ ms}$$

Use these values to compute the standard error for the difference in sample means. Compare this to the standard deviation you observed in the applet. In particular, have we over- or under- estimated the variability?

Standard error =

Comparison:

**Discussion:** This formula will underestimate the variability in the differences in sample means because with the random shuffling, we do not have “independent” samples. The overall sample mean for all 21 observations is constant, so if one group ends up with a higher mean, the other group must end up with a lower mean, leading to a larger difference in the group means. However, the real question is how the standardized statistic behaves.

(d) Calculate and interpret the two-sample *t*-statistic from Investigation 4.2 for the Sleep Deprivation data.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

(e) Use technology (e.g., [Comparing Groups \(Quantitative\)](#) applet) to examine a distribution of the *t* statistics for the Sleep Deprivation data. [In the applet, you can use the pull-down menu on the left to change the choice of Statistic to **t-statistic**. Then check the box to **Overlay t-distribution**.] Does the *t* probability distribution appear to be a reasonable model for the null distribution of the *t*-statistic?

The *t*-distribution comes to the rescue again and in most situations (moderate and large sample sizes) gives an adequate approximation to the exact randomization distribution. The next question is what to use for the degrees of freedom. We can use the same short-cut as in Section 2, the smaller of the two sample sizes minus one, or the “Welch-Satterwaite” approximation. [Keep in mind, this approximation is not always an integer.] For these data, technology reveals an approximate df of 17.56.

(f) Calculate the probability of obtaining a *t*-value of 2.69 or more, specifying 17.56 as the “Degrees of Freedom.”

- In R: use `iscamtpprob(2.69, df=17.56, "above")`

How does this probability compare to the exact p-value of 0.00696?

**Discussion:** Your analysis should reveal that the *t*-distribution provides a reasonable approximate model for the exact (standardized) randomization distribution. We can use the same standard error formula from Section 2; although this will underestimate the variability in our differences in group means, the *t*-statistic compensates for this in most situations and we find a reasonable approximation for the p-value. Will our randomization distribution always be well-modeled by a normal distribution (and therefore the test statistic by a *t*-distribution)? No, but we will consider this a reasonable approximation as long as either (similar to what you witnessed in Investigation 4.2):

- The data in both groups are symmetric and bell-shaped. When this is the case, we have evidence that the “treatment populations” are normally distributed. When this is true, the randomization distribution of the differences in group means will also follow a normal distribution.

OR

- The two sample sizes are large. Typically the sample sizes can be as small as 5, especially if the two groups have similar sample sizes and similar distribution shapes, but conventionally 20 is used as cut-off for how large each sample should be to use this approximation. Examine graphs of your data first. If the sample distributions are not symmetric or have unusual observations, you will want larger sample sizes.

In summary, we will often apply the two-sample *t*-procedures to both randomized experiments and to independent random samples. The distinction between these two sources of randomness will be most important in drawing your final conclusions (e.g., causation, generalizability).

- (g) The next question is what to use for the parameter in such a randomized experiment. Suggest symbols and an interpretation for the unknown parameter.

### Study Conclusions

The approximate p-value from the two-sample *t*-test ( $0.0076$ ,  $df = 17.56$ ) also provides very strong evidence that the observed difference between the two groups did arise by chance alone. Therefore, if we were to perform unlimited administrations of the same training and reaction time test under the exact same conditions, we have convincing evidence that the theoretical mean of all improvement scores that an individual would have under sleep deprivation is lower, three days later, than the theoretical mean of all improvement scores that an individual would have with unrestricted sleep (i.e.,  $\mu_{\text{sleepdeprived}} < \mu_{\text{unrestricted}}$ ). Though we still need to worry about what larger population these individuals represent.

### Practice Problem 4.5

Recall the study on children's television viewing habits from Practice Problems 2.5B and 4.3. One school incorporated a new curriculum, the other school did not.

- (a) Explain how the randomization in this study could be improved (in principle, even if difficult in practice, keeping in mind the goal of random assignment in general).

The following summary statistics pertain to the reports of television watching at the *conclusion* of the study:

Follow-up	Sample size	Sample mean	Sample SD
Control group	$n_1 = 103$	$\bar{x}_1 = 14.46$	$s_1 = 13.82$
Intervention group	$n_2 = 95$	$\bar{x}_2 = 8.80$	$s_2 = 10.41$

- (c) Are these sample means significantly different? Conduct a two-sample *t*-test of whether the long-run mean number of hours of television viewing per week is higher without the intervention than with the intervention after six months. State the null and alternative hypotheses, and report the test statistic and p-value. Also indicate whether you would reject the null hypothesis at the 0.05 level.

- (d) Summarize your conclusion from this test, including a discussion of causation and generalizability.

- (e) Explain why the nonnormality of these distributions does not hinder the validity of using this *t*-test procedure.

### Investigation 4.6: Ice Cream Serving Sizes

Researchers conducted a study in which they invited nutrition experts to an ice cream social (*Mindless Eating: Why we eat more than we think*, Wansink et al., 2006). These experts were randomly given either a 17- or a 34-ounce bowl. They were then invited to serve themselves ice cream. The suspicion was that even nutrition experts would tend to take more ice cream when given a larger bowl. Sample results for the actual volumes of ice cream taken (in ounces) are summarized in the following table:

	Sample size	Sample mean	Sample SD
<b>17-ounce bowl</b>	20	4.38	2.05
<b>34-ounce bowl</b>	17	5.81	2.26

- (a) Identify the explanatory and response variables in this study.

Explanatory:

Response:

- (b) Specify the null and alternative hypotheses for this research study, in symbols and in words.

- (c) Would it be possible for you to carry out a randomization test using simulation with the information presented here? Explain.

- (d) Use technology to carry out a two-sample  $t$ -test to determine whether individuals given a 34-oz bowl have a tendency to serve themselves more ice cream than individuals given a 17-oz bowl. Be sure to report your test statistic and p-value. Also state a conclusion that addresses the statistical significance of the data, whether you are willing to draw a cause and effect conclusion, and what population you are willing to generalize the results to.

- (e) Now that the significance test has provided moderately strong evidence that nutrition experts given a larger bowl tend to take larger servings, what might a natural follow-up question be?

When a test of significance indicates that the difference between two group means is statistically significant, it makes sense to estimate the magnitude of that difference in the population or treatment means with a confidence interval. We will use the same two-sample  $t$ -confidence interval method from Investigation 4.5 to estimate a confidence interval.

(f) Use technology (see the Technology Detour to Investigation 4.2) to determine a 95% confidence interval. (Recall that you will need to change the alternative to “not equal” in Minitab.) Report the endpoints of the interval as well as the margin-of-error. Interpret the interval, being sure to specify clearly the parameter being estimated.

Margin-of-error:

Interval:

Interpretation of confidence interval:

(g) Is your confidence interval calculation consistent with what you found with the test of significance? Explain.

### Study Conclusions

In this study, we find that the observed difference in the average serving size taken by nutrition experts using a 34-oz bowl is significantly larger than for those using the 17-oz bowl ( $p$ -value  $\approx 0.027$ ). In fact, we can say that we are 95% confident that if we were to repeat these identical conditions indefinitely, those using the larger bowls would take on average up to 2.8555 more ounces of ice cream. However, the two-sided  $p$ -value is larger than 0.05, so our 95% CI does include zero, and it’s plausible that those using the smaller bowls will take on average up to 0.0255 fewer ounces of ice cream in the long-run.

### Practice Problem 4.6

- Calculate and interpret a 99% confidence interval for these data.
- How will this interval change if we lower the confidence level to 90%? Explain, without performing a calculation.
- How will the width and/or midpoint of this interval change if both sample standard deviations are doubled? Be as specific as possible, and explain without determining the new confidence interval.
- How will the width and/or midpoint of this interval change if both sample sizes are doubled? Be as specific as possible, and explain without determining the new confidence interval.
- How will the width and/or midpoint of this interval change if both sample means are increased by two ounces? Be as specific as possible, and explain without determining the new confidence interval.

### Investigation 4.7: Cloud Seeding

Our lives depend on rainfall. Consequently, scientists have long investigated whether humans can intervene and, as needed, help nature produce more rainfall. In one study, researchers in southern Florida explored whether injecting silver iodide into cumulus clouds would lead to increased rainfall. On each of 52 days that were judged to be suitable for cloud seeding, a target cloud was identified and a plane flew through the target cloud in order to seed it. Randomization was used to determine whether or not to load a seeding mechanism and seed the target cloud with silver iodide on that day. Radar was used to measure the volume of rainfall from the selected cloud during the next 24 hours. The results below and in [CloudSeeding.txt](#) (from Simpson, Alsen, and Eden, 1975) measure rainfall in volume units of acre-feet, “height” of rain across one acre.

Unseeded:

1.0	4.9	4.9	11.5	17.3	21.7	24.4	26.1	26.3	28.6	29.0	36.6	41.1
47.3	68.5	81.2	87.0	95.0	147.8	163.0	244.3	321.2	345.5	372.4	830.1	1202.6

Seeded:

4.1	7.7	17.5	31.4	32.7	40.6	92.4	115.3	118.3	119.0	129.6	198.6	200.7
242.5	255.0	274.7	274.7	302.8	334.1	430.0	489.1	703.4	978.0	1656.0	1697.8	2745.6

(a) Identify the observational/experimental units, explanatory variable, and response variable.

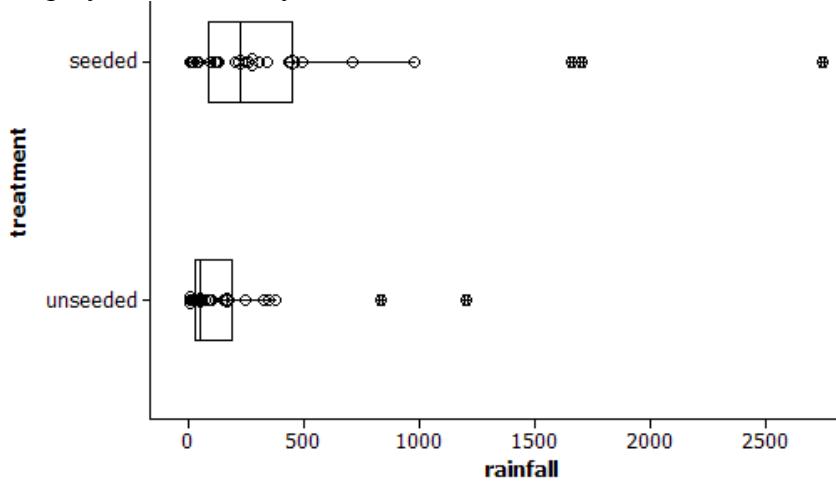
Observational/experimental units:

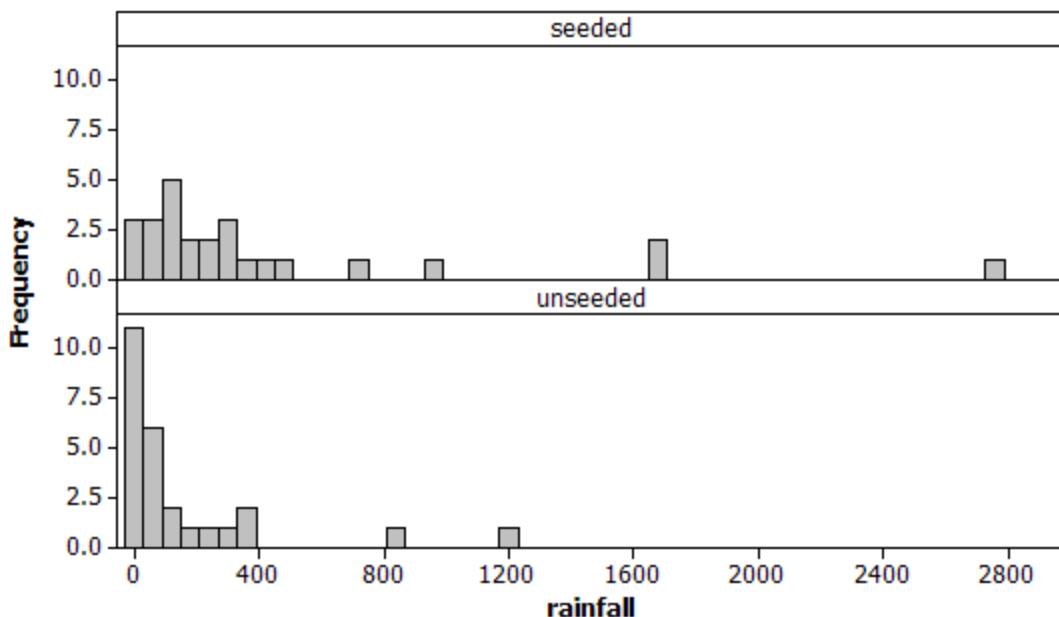
Explanatory:

Response:

(b) Is this an experiment or an observational study? Explain.

Below are graphical displays and summary statistics for these two distributions





Variable	treatment	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
rainfall	seeded	26	442	651	4	79	222	445	2746
	unseeded	26	164.6	278.4	1.0	23.7	44.2	183.3	1202.6

(c) Based on the graphs and summary statistics, is there preliminary evidence that cloud seeding is effective? (Include a calculation of the difference in the group means and in the group medians.)

(d) Would it be reasonable to apply a two-sample  $t$ -test to assess the statistical significance of the difference in the sample means? Explain.

(e) Regardless of your answer to (d), carry out a two-sample  $t$ -test and a confidence interval to determine the strength of evidence of higher average rainfall with the seeded clouds.

(f) What other approaches have you learned that could be applied here instead of a two-sample  $t$ -test to assess the statistical significance of the difference in group means?

One approach to compensate for the skewness and outliers in these rainfall distributions is to analyze the difference in *medians* instead of the difference in means. Fortunately, an advantage of the randomization test is that it applies just as well, and just as easily, to comparing medians.

(g) Modify the previous R or Minitab code from the Technology Detour in Investigation 4.4 (remember to load/attach the [CloudSeeding.txt](#) data, to match the sample sizes in the actual study, and to consider the difference in medians as the statistic), or use the [Comparing Groups \(Quantitative\)](#) applet, to calculate an empirical p-value for the difference in medians for this study. [Hint: What is the observed difference?] Report the p-value and your conclusion about the effect of cloud seeding on median rainfall amounts.

(h) Would it be reasonable to model the randomization distribution for the difference in medians with a normal or *t*-distribution? Explain.

Although we can certainly use simulation this way, it is not feasible to find the “exact” randomization distribution for a data set of this size, and we also notice that the distribution of the difference in medians has less “smoothness” than the distribution of the differences in means, as certain combinations values are just not possible from a discrete distribution.

Another approach you saw earlier is to *transform* the data first.

(i) Determine the (natural) logged rainfall values.

- In R, recall > `lnrainfall = log(rainfall)`

(j) Create dotplots [Hint: `iscamdotplot` or Graph > Dotplot] of *ln rainfall* amounts by treatment.

Would you consider the shapes of the distributions to be more symmetric than with the original data?

Would you consider the variability to now be more similar between the two groups?

Is there still evidence of a positive shift in (ln) rainfall amounts for the seeded clouds?

(k) Would you consider it more reasonable to apply two-sample *t*-procedures to the transformed rainfall amounts than the original data? Explain.

(l) Calculate the one-sided p-value and the (two-sided) 95% confidence interval based on these data. [Hint: Pay attention to the direction of subtraction by the technology.]

- (m) How does this p-value compare to what you found from simulating a randomization distribution for the difference in medians? What conclusion would you draw from this p-value?

**Discussion:** It is fairly straight-forward to “back-transform” our transformation by exponentiating the endpoints of the interval. But we will again focus on the median rather than the mean as the transformed parameter. Also exponentiating an interval for a difference really gives us a ratio in the original scaling [ $10^{\mu_1 - \mu_2} = 10^{\mu_1}/10^{\mu_2}$ ]. In other words, if group 1 is 2 points higher than group 2 on the log scale ( $\mu_1 = \mu_2 + 2$ ), then group 1 is  $10^{\mu_2+2} = 10^{\mu_2}(100)$  or 100 times higher in the original scale. Therefore, once we apply the back-transformation, we really have an interval of plausible values for the *multiplicative* change in the *median* of the response variable.

- (n) Perform the back-transformations and interpret the new confidence interval in terms of a ratio of treatment medians.

### Study Conclusions

The rainfall distributions reveal that the rainfall amounts from the seeded clouds have a larger mean and larger variability than from unseeded clouds, and both groups have very right-skewed distributions of rainfall with several outliers. A two-sample *t*-test may not be valid for these data so we could consider a randomization test on the medians, but then we don’t have a convenient way to find a confidence interval. This is a situation where a ln-transformation is likely to be helpful in rescaling the data so that the *t*-procedures are more valid. Examining graphs of the transformed data confirm that the variability is now more similar and both shapes are symmetric. The two-sample *t*-test applied to the ln-transformed data can thus be used as an approximation to the randomization test. Also, the confidence interval for the difference in the ln rainfall can be back-transformed to provide a statement about the multiplicative treatment effect of cloud seeding: We are 95% confident that the median volume of rainfall on days when clouds are seeded is 1.3 to 7.7 times larger when clouds are seeded as opposed to when they are not seeded. Because random assignment was used to assign the clouds to the seeding conditions, it is safe to interpret this as evidence that the seeding causes larger rainfall amounts on average.

Although the transformation again gives us a comparison of medians, which is often preferable with skewed data and/or outliers, keep in mind, sometimes we really prefer the parameter to be a mean than a median, such as when you want to focus on the “total” amount (population mean  $\times$  population size).

### Practice Problem 4.7

Compare these two analyses to a two-sample *t*-test and a two-sample *t*-confidence interval on the original data.

## SECTION 4: MATCHED PAIRS DESIGNS

In this section, you will again compare two sets of quantitative observations but with one key difference in how the data were collected. You will see advantages to collecting data with a “paired” design, and then you will investigate how to make the corresponding changes in the appropriate analyses.

### Investigation 4.8: Chip Melting Times

Suppose you want to compare the melting times of semisweet chocolate chips to peanut butter chips using the students in your class. To melt the chips, one chip will be placed on each student’s tongue and then held against the roof of his or her mouth. The time until the chip is completely melted, without any “encouragement” by the student, will be recorded (in seconds).

#### Design

(a) Describe a completely randomized experiment for conducting this study. Identify the experimental units and variables of interest. Classify the variables as explanatory and response, as well as quantitative or categorical.

Design:

Experimental units:

Explanatory variable:

Type:

Response variable:

Type:

(b) Carry out this study in your class, submitting the data as instructed. Be sure to record which type of chip you had and the melting time (in *seconds*). Your instructor will then make the class results available to you.

#### Descriptive Statistics

(c) Load the data into R or Minitab. Examine stacked dotplots (`iscamdotplot`) of the two sets of melting times as well descriptive statistics (e.g., **Stat > Basic Statistics > Display Descriptive Statistics** or `iscamsummary()` possibly using `na.omit`). Comment on how the shape, center, and variability of the two distributions compare.

Chocolate mean:

Peanut butter mean:

Chocolate SD:

Peanut butter SD:

Comparison:

**Statistical Inference**

(d) Now carry out a two-sample  $t$ -test to analyze these data. Be sure to state the null and alternative hypotheses. Also comment on whether or not you believe the two-sample  $t$ -test is valid with these data.

(e) Suppose there truly is a difference in the average melting times by college students between these two types of chips. Explain why you may not be able to detect this difference with this study.

**Alternative Study Design**

(f) Is there a way to modify the experimental design that will give you a better chance of detecting a difference between melting times of these types of chips if one exists? Be sure to explain why you believe this new design will be advantageous in detecting a difference between the two chips.

(g) Carry out the second phase of this study and again record the results for your instructor. Be sure to enter the requested information correctly (matching the time with the chip type no matter which you did first).

(h) Is this study still an experiment? Explain. Was randomness used? For what purpose? What are the observational/experimental units? What variable will we analyze for each observational/ experimental unit?

Type of study:

Randomness/Why:

Observational/experimental units:

Variable:

Let's again look at comparative dotplots and descriptive statistics for the chocolate times and the peanut butter times. This time when you copy in the data, the chocolate and peanut butter times are in two different columns ("unstacked") so we will create the graphs a little bit differently.

### Technology Detour – Comparative Dotplots (with data in separate columns)

In R (assuming dataset is called "chiptimes" with variables "choc" and "pb")

```
> StackedData = stack(chiptimes[, c("choc", "pb")])
> names(StackedData) = c("meltttime", "chiptype")
> attach(StackedData)
> iscamdotplot(meltttime, chiptype, names=c("times", "chip type"))
```

- (i) Also calculate descriptive statistics (mean and standard deviation) of the melting times as before (using the stacked data in R, putting both columns in the Variables box in Minitab).

Chocolate mean:

Peanut butter mean:

Chocolate SD:

Peanut butter SD:

How do the distributions of melting times, as revealed by these graphs and descriptive statistics, compare to your earlier analysis (from the randomized comparative experiment)?

- (j) Would it be valid to carry out a two-sample *t*-test to compare the average chocolate chip times to the average peanut butter chip times for these new data (matched pairs experiment)? Explain.

When the data are *paired* (e.g., repeat observations on the same individual) we should not treat them as *independent* samples as you considered doing in (i) and (j). This ignores the information that two measurements were taken for each person (we couldn't mix up the values in the second column without altering the information in the data). Instead you can analyze the *differences* in the times per person.

- (k) Calculate the differences between the chocolate chip times and the peanut butter chip times for each person (chocolate – peanut butter). [In R, you may have to attach "chiptimes" again.] Examine and describe a dotplot and descriptive statistics for these differences. Do they support a tendency for one chip to melt more slowly than the other? How are you deciding?

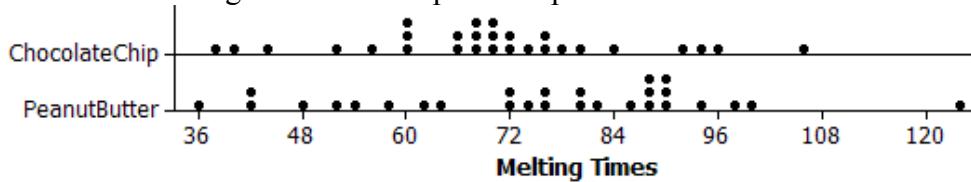
Mean difference:

Standard deviation of differences:

(l) Does the pairing in the study design appear to have been helpful here? To decide, compare the standard deviation of the differences to the individual peanut butter and chocolate chip standard deviations.

(m) Do you believe the evidence for a difference between chocolate chips and peanut butter chips is stronger, weaker, or similar to the strength of evidence from the completely randomized design? In particular, do you think a p-value we might get from these differences will be larger, smaller, or similar to what you would find from a two-sample  $t$ -test? Explain, being sure to consider factors that affect the size of a p-value when comparing a quantitative response variable between two groups.

(n) Consider the following results from a paired experiment.



Suggest a scenario where the paired data would be pretty convincing that the chocolate chip melting time is shorter, despite the substantial overlap between the two distributions. Then suggest a scenario where the data are paired but a significant difference is not found in the melting times.

**Discussion:** Typically when we collect data on chip melting data from our students, it is difficult to find a statistically significant difference between the melting times of the two types of chips because there is so much person-to-person variability in the melting times (and therefore a lot of overlap in the two distributions). One way to control for this person-to-person variability is to have each person melt both types of chips. Now any differences we see between times is more directly related to the chip type. (The ultimate goal is to have chip type be the only difference between the two measurements.) If it is really the case that individual people have an effect on melting time, perhaps because of different body temperatures in their mouths or different melting techniques, those differences will now apply to both

types of chips, and we can instead focus on the differences between the chips times when they are melted under identical conditions. Consequently, there should be much less variability in the *differences* than in the individual melting times. This decrease in variability will in turn increase our “power” to detect any underlying differences between the two chip types (even without technically increasing the sample size in the study). If the variability in the differences is not smaller, then this would be a case where the pairing didn’t really help (though also didn’t really hurt other than the additional time requirement to collect the second set of observations). Do keep in mind that with paired designs, randomization is still very important. In a matched-pairs experiment, you should randomize which of the two treatments each subject receives first. This will balance out any confounding variables related to time (e.g., learning curve) between the two chip types.

### Practice Problem 4.8A

Suppose that a baseball manager wants to study whether a player can run from second base to home plate more quickly by taking a wide angle around third base or a narrow angle. Forty players are available to use as subjects in an experiment.

- (a) Suggest a better experimental design than randomly assigning 20 players to run each angle.
- (b) What is the primary advantage of the matched-pairs design in this study?

### Practice Problem 4.8B

For each of the following research study designs, indicate whether the data collection plan will result in two independent samples (completely randomized design) or dependent samples (matched-pairs design).

- (a) A farmer wants to see whether referring to cows by name increases their milk production. He selects half of his cows at random, gives them names, and frequently calls them by name. The other half of his cows he does not call by name. Then he measures the milk production of each cow over a one-week period.
- (b) A farmer wants to know whether hand-milking or machine-milking tends to produce more milk from cows. He examines records of how much milk the cows have produced in the past, and order them from most to least productive. For the top two milk producers, randomly assign one to hand-milking and the other to machine-milking. Do the same for the next two and the next two and so on.
- (c) You wonder whether students at your school tend to drive newer cars than faculty at your school. You take a random sample of 20 students and a random sample of 20 faculty members, and ask each person how old their car is.
- (d) To investigate whether knee surgery is effective, you randomly assign half of the subjects to receive knee surgery and the other half to undergo a “placebo” operation.
- (e) To investigate the effectiveness of an online language study program, participants were assigned to enroll in a six-week summer session, after which their language skills were assessed, and then to spend six-weeks using an online program (Duolingo), after which their language skills were assessed.

**Investigation 4.9: Chip Melting Times (cont.)**

In the previous investigation, you began to explore the advantages of a matched pairs experimental design over a completely randomized design. Now you will focus on how to analyze data that has been collected from a paired design. As always, you will first consider how to simulate a distribution of possible statistics assuming the null hypothesis and then you will consider the appropriateness of a mathematical model.

- (a) For the matched-pairs design that you collected (where each of you tried both types of chips), consider the *differences in melting times* as the response variable. Is this variable quantitative or categorical? What are the observational units? Is there also an explanatory variable?
- (b) Define an appropriate parameter to investigate whether one type of chip tends to melt slower than the other and state a null and an alternative hypothesis about this parameter.

**Simulation**

(c) Outline how you could use a coin to simulate a randomization test to compare the two sets of measurements to assess how unusual it is for the average difference in times to be at least this different from each other just by chance. Keep in mind that you want the simulation to mimic the randomization process used in the study design, assuming the type of chip does not affect the melting times.

- (d) Describe the components of the code (pseudo-code) you could use to carry out this simulation using a computer.

(e) Copy and paste the original raw data into the [Matched Pairs Randomization](#) applet:

- View the original data in a spreadsheet, with one column for the peanut butter times and a second column for the chocolate chip times (each row is one person). You can also include an initial column of identifiers (e.g., student IDs or initials).
- In the applet, press **Clear** to empty the data window and then paste in the three columns (with header information, column of IDs is optional but needs to be listed first), and press **Use Data**. The dotplots should then show both sets of data, connecting the paired observations, and their differences.
- Check the **Randomize** box and press **Randomize**. For each pair, the applet will virtually “flip a coin” and if it lands heads, the two observations for that person will change positions. The new dotplots and the new set of differences for these rearranged values will be displayed. The mean of these differences will appear in the bottom dotplot.
- Uncheck **Animate**.
- Press **Randomize** four more times to get a sense of the variability in the results from repetition to repetition.
- Change the number of randomizations from 1 to 995 (for a total of 1000) and press **Randomize**.

(f) Explain what distribution is being displayed in the bottom dotplot.

(g) Where is the distribution of the average differences centered?

How surprising does our observed value for the mean difference appear to be, under the simulation’s assumption that type of chip does not affect melting time?

(h) Count the number of simulated Average Differences that are more extreme than what we observed and report the empirical p-value. [Hint: Be sure to consider the alternative hypothesis when deciding what to consider as “more extreme.”] What conclusion will you come to based on this p-value? Can you draw a cause and effect conclusion? For what population?

Paired Data:			
id	narrow	wide	
1	5.5	5.55	
2	5.7	5.75	
3	5.6	5.5	
4	5.5	5.4	
5	5.85	5.7	
6	5.55	5.6	
7	5.4	5.35	
8	5.5	5.35	
9	5.15	5	



## Mathematical Model

(i) Does the randomization distribution appear that it would be well modeled by a normal distribution? If you select the ***t*-statistic** radio button, do the standardized statistics appear well-modelled by a *t*-distribution?

**Definitions:** A [paired \*t\*-test](#) standardizes the mean of the *differences* from a matched-pairs design.

$$t = (\bar{x}_d - 0) / (s_d / \sqrt{n_d})$$

where  $\bar{x}_d$  is the sample mean of differences and  $s_d$  is the sample standard deviation of the differences. The test statistic above assumes the hypothesized difference is zero, but this can be changed.

**Technical conditions:** When the distribution of differences is normally distributed or the sample size is large (e.g.,  $n \geq 30$  pairs of observations), this *t*-statistic is well modeled by a *t*-distribution with  $n - 1$  degrees of freedom.

A [paired \*t\*-confidence interval](#) for  $\mu_d$  has the form  $\bar{x}_d \pm t_{n-1}^*(s_d / \sqrt{n_d})$

**Note:** These are a special case of the *one-sample *t*-procedures* that can be applied to a single sample of quantitative data (see Investigation 2.5). In this case, the sample consists of the differences.

(j) Calculate and interpret the value of this test statistic, by hand, based on the summary statistics.

After calculating this test statistic, you can use the *t*-distribution to determine the p-value. You can also use technology to carry out the significance test calculations, as well as to determine a 95% confidence interval for the mean difference in the population.

## Technology Detour – Paired *t*-tests

**In R:** [you can use the \*t.test\* command as before but specify the data are paired, e.g.,](#)

```
> t.test(choc, pb, alternative="greater", conf.level = .95
+ paired=TRUE)
OR
> t.test(melttime~chiptype, alt="greater", paired=TRUE)
```

(k) Report the test statistic and p-value for this paired *t*-test. How does the p-value compare to the empirical p-value from the simulated randomization test?

- (l) We suspect the p-value for this analysis is smaller than the p-value for the two-sample *t*-test conducted for the first (completely randomized design) dataset. Explain why the *power* of the paired study design is expected to be greater for the paired study than for the completely randomized design.
- (m) Report and interpret the 95% confidence interval calculated using technology.
- (n) How does this confidence interval compare to the one based on the unpaired (completely randomized design) data study?

**Discussion:** Two compare two means, a more powerful study design than randomly assigning subjects to two groups, if possible, is to pair the observations. This pairing accounts for the variability from observational unit to observational unit and allows for a more direct comparison between the two conditions. For example, if two measurements are taken on each experimental unit, there should not be any other systematic differences between the measurements other than the treatment effect and random chance. Randomizing will still be important in determining the *order* of the two treatments, thereby eliminating order as a potential confounding variable. By accounting for the unit-to-unit variability, this should increase the power of the test of significance, making it easier to detect a difference between the two conditions if one really exists. To analyze such data, perform a matched-pairs randomization test or a (one sample) *t*-test on the *differences*.

### Study Conclusions

After carrying out a paired experiment comparing chocolate chip melting times to peanut butter chip melting times, you probably saw some evidence of longer melting times for the peanut butter chips. Carrying out a paired *t*-test on the differences, you can assess the statistical significance of this observed time difference. You can also calculate a confidence interval for the *mean difference* (rather than the difference in means) to estimate the size of the underlying treatment effect, on average, for the overall process or population. You can judge whether the pairing was beneficial by seeing whether the variability in the differences is smaller than the variability of the melting times.

### Practice Problem 4.9

Scientists have long been interested in whether there are physiological indicators of diseases such as schizophrenia. In a 1990 study by Suddath et. al., reported in Ramsey and Schafer (2002), researchers used magnetic resonance imaging to measure the volumes of various regions of the brain for a sample of 15 monozygotic twins, where one twin was affected by schizophrenia and other not (“unaffected”). The

twins were found in a search through the United States and Canada, the ages ranged from 25 to 44 years, with 8 male and 7 female pairs. The data (in cubic centimeters) for the left hippocampus region of the brain are in [hippocampus.txt](#). The primary research question is whether the data provide evidence of a difference in hippocampus volumes between those affected by schizophrenia and those unaffected.

- (a) Calculate the difference in hippocampus volumes for each pair of twins (unaffected – affected).
- (b) Calculate and interpret a 95% confidence interval for the mean volume difference using the paired *t*-interval. Also comment on the validity of this procedure.
- (c) Based on this confidence interval, is there statistically significant evidence that the mean difference in left hippocampus volumes is different from zero? Explain.

### Investigation 4.10: Comparison Shopping

In 1999, student group at Cal Poly carried out a study to compare the prices at two different local grocery stores. The inventory list for Scolari's (a privately-owned local store) was broken into 32 sheets, each with 30 items. A number between 01 and 30 was randomly generated, 17, and the 17<sup>th</sup> item on each sheet was selected. Each student was then responsible for obtaining the price of that item at both Scolari's and Lucky's (which advertises itself as a discount grocery store). If the exact item was not available at both stores, the item was adjusted slightly (size or brand name) so that the identical item could be priced at both stores. Students at Cal Poly gathered two prices for a set of 29 items. The data are available in [shopping99.txt](#).

- (a) Identify the observational units, response variable, population, and sample in this study. Which store do you suspect will have lower prices?

Observational units:

Response variable:

Population:

Sample:

Initial conjecture:

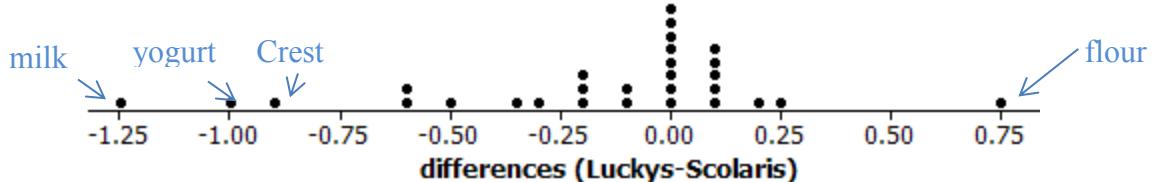
- (b) Was this an experimental study or an observational study? Was this a simple random sample? Was it a probability sample? What is the advantage of this procedure over a simple random sample? Are there any disadvantages in selecting the sample this way?

- (c) Did these students use independent samples or a paired design to collect their data?

- (d) Define the parameter of interest and state the null and alternative hypotheses corresponding to your initial conjecture.

Descriptive statistics for the items' prices (in dollars) at each store, and for the differences in prices, as well as a dotplot of the differences are shown here:

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Lucky's	31	2.408	1.660	0.490	1.090	1.990	2.790	6.990
Scolaris	29	2.593	1.742	0.500	1.020	2.300	3.585	6.790
diffs	29	-0.1566	0.4083	-1.2300	-0.3250	0.0000	0.1000	0.7600



After further investigation on the outliers, only the milk was found to be recorded incorrectly, two different sizes, and it was removed from the data set.

- (e) Does the pairing appear to have been useful here? Discuss both why you suspect pairing will be beneficial in this context and any evidence in the above output.

- (f) Would applying a paired *t*-test appear to be appropriate for these data? Explain.

- (g) Below is the output from R of a paired *t*-test and a 90% confidence interval:

```
data: Shopping$Lucky's and Shopping$Scolaris
t = -1.7435, df = 27, p-value = 0.04631
alternative hypothesis: true difference in means is less than 0
sample estimates:
mean of the differences           -0.1182143

90 percent confidence interval:
-0.233700560 -0.002728012
```

What conclusions can you draw from this output? Be sure to address significance, causation, generalizability, and confidence.

- (h) Calculate (by hand) a 90% *prediction* interval (Investigation 2.6) based on this sample. Include a one-sentence summary of what this interval says.

- (i) How does the width of this prediction interval in (h) compare to the width of the confidence interval for the population mean price difference in (g)? Explain why this makes sense. [Hint: Which is harder to predict, the mean or the next observation?]
- (j) Which interval, the prediction interval or the confidence interval, do you find more useful here? Explain.

### Study Conclusions

An important first step in data analysis is always to explore your data! With these data, we found some unusual observations and upon further investigation realized that one data value had been recorded in error (e.g., milk). In this case, we would be justified in removing this observation from the data file (we did not have any justification for removing the other outliers). After cleaning the data, we found that the price differences were slightly skewed to the left with a few outliers (flour, toothpaste, and frozen yogurt). The average price difference (Lucky's – Scolari's) was  $-\$0.118$ , with a standard deviation of  $\$0.359$ . The median price difference was  $\$0$ . So though the sample did not show a strong tendency for one store to have lower prices, the sample mean difference was in the conjectured direction. A one-sided *paired t-test* found that the mean price difference between these two stores was significantly less than  $0$  ( $t$ -value =  $-1.74$ ,  $p$ -value =  $0.046$ ) at the  $10\%$  level of significance, and even (barely) at the  $5\%$  level. A  $90\%$  confidence interval for the mean price difference was  $(-0.234, -0.003)$ . We are  $90\%$  confident that, on average, items at Scolari's cost between  $0.3$  cents and  $23$  cents more than items at Lucky's. This seems like a small savings but could become practically significant for a very large shopping trip. (Note, the average savings is not the same as the savings we would expect on an individual item.) In fact, we are  $90\%$  confident that an individual item will be anywhere from  $74$  cents more expensive at Scolari's to  $50$  cents more expensive at Lucky's. We feel comfortable generalizing these conclusions to the population of all products common to the two stores because the data were randomly selected using a probability method (systematic random sampling).

### Practice Problem 4.10

- (a) Reconsider the hippocampus values from Practice Problem 4.9 ([hippocampus.txt](#)). Calculate and interpret a  $95\%$  prediction interval using these data.
- (b) Do you believe this is a valid procedure for these data? Explain.

**Summary of Procedures for Paired Differences****Test of  $H_0: \mu_d = \mu_0$** 

- Randomization test (randomizing the sign of the difference)
- Paired *t*-test (Valid with  $n \geq 30$  or normal population of differences)

$$\text{Test statistic: } t_0 = (\bar{x}_d - \mu_0) / (s_d / \sqrt{n})$$

Degrees of freedom =  $n - 1$

***t*-Confidence interval for  $\mu_d$ :**  $\bar{x}_d \pm t^*_{n-1} \times s_d / \sqrt{n}$

Valid with  $n \geq 30$  or normal population of differences

**In Minitab:** Stat > Basic Statistics > Paired t

**In R:** `t.test(..., paired=TRUE)`

*You can also apply one-sample *t*-procedures to the differences.*

### Investigation 4.11: Smoke Alarms

A study published in the journal *Pediatrics* (Smith et al., 2006) addressed the important issue of how to awaken children during a house fire so they can escape safely. Researchers worked with a volunteer sample of 24 healthy children aged 6-12 by training them to perform a simulated self-rescue escape procedure when they heard an alarm. Researchers then compared the children's reactions to two kinds of alarms: a conventional smoke alarm and a personalized recording of the mother's voice saying the child's name and urging him or her to wake up. All 24 children were exposed to both kinds of alarms, with the order determined randomly.

- (a) Identify the observational units, explanatory variable, and response variable. Are the variables quantitative or categorical? Is this an observational study or an experiment? If this is an experiment, how was randomization used?

Observational units:

Explanatory variable:

Type:

Response variable:

Type:

Type of Study:

Randomness:

It turned out that one child did not wake up to either alarm, 14 woke up to both alarms, and 9 woke up to the mother's voice but not the conventional alarm. We could create a two-way table to summarize these data, again using the explanatory variable as the columns.

	Conventional alarm	Mother's voice	Total
Awakened			
Did not awaken			
Total			

- (b) Would this be an appropriate way to display and then analyze these data? Explain how you can tell.

- (c) To maintain the pairing in the data, we could instead use a table as follow. Use the above information to complete the table.

	Awakened to conventional alarm	Did not awaken to conventional alarm	Row total
Awakened to mother's voice			
Did not awaken to mother's voice			
Column total			

- (d) At first glance, without doing any formal analysis, does it appear that the study provides evidence that the alarm based on the mother's voice is more effective at waking children than the conventional alarm? Explain.

In this study, every child experienced both the conventional tone alarm and the mother's voice alarm. Randomization was still used for each child, but only to determine which kind of alarm was experienced first. This is again a matched pairs design, but with a *categorical* response. So the analysis question is how do we focus on the differences in the reaction to the two alarms for each child?

Because of the paired nature of the experimental design, the data could be entered into a computer package as follows. Notice that, in keeping with convention, the observational units (children) are in rows and the variables are in columns:

<b>Child #</b>	<b>Response to mother voice</b>	<b>Response to conventional</b>	<b>Child #</b>	<b>Response to mother voice</b>	<b>Response to conventional</b>
1	wake	wake	13	wake	wake
2	wake	wake	14	wake	wake
3	wake	wake	15	wake	not
4	wake	wake	16	wake	not
5	wake	wake	17	wake	not
6	wake	wake	18	wake	not
7	wake	wake	19	wake	not
8	wake	wake	20	wake	not
9	wake	wake	21	wake	not
10	wake	wake	22	wake	not
11	wake	wake	23	wake	not
12	wake	wake	24	not	not

As always, the big question is: How surprising would the observed experimental result be, under the null model that there's really no difference in success rates between the two kinds of alarms? Also as always, we can address this question by re-randomizing a large number of times to see what's typical and what's unusual.

- (e) Outline how you could conduct a simulation analysis to address this research question.

Remember that every child was subjected to both treatments, and the randomization determined which kind of alarm the child experienced first. So, we will re-randomize for each subject, assuming that their two reactions (wake or not) would have been the same, but flipping a coin to decide which reaction goes with the conventional tone and which with the mother's voice.

**Short-Cut:** Notice that we can take a time-saving shortcut before we start this. Children #1-14 and #24 had the same reaction for both treatments, so there's no need to re-randomize for those children. We will focus on children #15-23, who had different reactions (one wake and one not) to the two treatments. For each of these nine children, we will flip a coin to determine whether the "wake" reaction goes with the conventional alarm or with the mother's voice alarm.

(f) Flip a coin nine times, once for each of children #15-23. If the coin lands heads, the wake reaction stays with the mother's voice alarm and the not wake with the conventional alarm. But if it lands tails, the wake reaction moves to the conventional alarm and the not wake moves to the mother's voice alarm.

Child #	Response to mother voice	Response to conventional	Re-randomized mother voice	Re-randomized conventional voice
1	wake	wake	wake	wake
2	wake	wake	wake	wake
3	wake	wake	wake	wake
4	wake	wake	wake	wake
5	wake	wake	wake	wake
6	wake	wake	wake	wake
7	wake	wake	wake	wake
8	wake	wake	wake	wake
9	wake	wake	wake	wake
10	wake	wake	wake	wake
11	wake	wake	wake	wake
12	wake	wake	wake	wake
13	wake	wake	wake	wake
14	wake	wake	wake	wake
15	wake	not		
16	wake	not		
17	wake	not		
18	wake	not		
19	wake	not		
20	wake	not		
21	wake	not		
22	wake	not		
23	wake	not		
24	not	not	not	not

Create the table corresponding to your results:

	Awakened to conventional alarm	Did not awaken to conventional alarm	Row total
<b>Awakened to mother's voice</b>	14		
<b>Did not awaken to mother's voice</b>		1	
<b>Column total</b>			24

(g) Is your simulated table as extreme (favoring the mother's voice alarm) as the actual experimental data? How are you deciding?

(h) How many heads are needed to obtain a result at least as extreme as the actual experimental result? In other words, among the nine children who reacted differently to the two treatments, how many woke to the mother's voice but not to the conventional alarm?

(i) We want to look at a much larger number of re-randomizations, so we will turn to the [One Proportion Inference](#) applet. Specify 0.5 as the probability of success and 9 as the sample size. Set the **Number of samples** to some large number like 1000. Press the **Draw Samples** button and describe the resulting distribution. How often did we observe 9 or more successes?

(j) Does this approximate p-value suggest that the observed data (namely, all 9 who had different reactions to the two alarms were awoken by the mother's voice) would be very unlikely to occur by chance alone, if there were no positive effect of the mother's voice? Explain.

(k) Use the binomial distribution to calculate the exact p-value for this test. (You can press **Exact Binomial** in the applet). Be sure to specify the parameter values and also indicate how you perform this calculation.

This test, using the binomial distribution to look at the distribution number of successes in the off-diagonal cells, is called **McNemar's Test**.

- (l) Summarize your conclusion for this smoke alarm study, and explain the reasoning process by which it follows.

### Study Conclusions

In this study, nine children reacted differently to the two alarms. If there were no effect from the type of alarm, we would expect half the children to respond to the mother's voice but not the conventional alarm and the other half of the children to respond to the conventional alarm but not the mother's voice. Instead, all 9 of the one-alarm children responded to the mother's voice but not the conventional alarm. This provides evidence that the mother's voice is more effective than the conventional alarm. To see whether this difference is statistically significant, we can calculate the binomial probability (assuming  $n = 9$  and  $\pi = 0.5$ ) to determine the exact p-value  $P(X \geq 9) = 0.00195$ . This provides strong evidence ( $p\text{-value} < 0.05$ ) that the observed result did not arise from the randomness in the process alone (which alarm was tested first). Because this was a randomized matched-pairs experiment, we will attribute this difference to the type of alarm. Keep in mind that this was a sample of volunteers and may not represent a larger population of children.

### Practice Problem 4.11

Another aspect of this same study considered not just whether the child woke up but whether he/she successfully escaped from the house within 5 minutes of the alarm sounding. The article reports that 20 children escaped when they heard the mother's voice, and only 9 escaped when they heard the conventional tone.

- (a) Describe what additional information you need before you can analyze the data as you did above.  
 (b) Two children did not escape to either kind of alarm. Use this information to complete the following table:

	Escaped to conventional alarm	Did not escape to conventional alarm	Total
Escaped to mother's voice			20
Did not escape to mother's voice			
Total	9		24

- (c) Conduct a simulation analysis (with the One Proportion Inference applet) of these data. Be sure to describe how you use the applet, and report an approximate p-value.

- (d) Use the binomial distribution to calculate the exact p-value for this test.  
 (e) Summarize your conclusion from this "escaping within 5 minutes" aspect of the study, and explain the reasoning process behind your conclusion.

**Example 4.1: Age Discrimination?**

Try these questions yourself before you use the solutions following to check your answers.

In Investigation 4.1, you considered the case of Robert Martin and whether the sample of 10 employees in his department provided evidence of age discrimination. Suppose we decide to focus on the long-run difference in mean ages for those laid-off and those retained for the decision making process used by this company.

(a) State the null and alternative hypotheses, in symbols, for this study. [Hint: Define your symbols.]

(b) Recall the observed value of the sample statistic.

(c) Simulate a randomization test for these data and state your conclusion at the 0.01 level of significance.

(d) Carry out a two-sample *t*-test for these data and hypotheses, and state your conclusion at the 0.01 level of significance.

(e) Do these analyses reach the same conclusion? If not, which analysis should be used? Explain.

**Analysis**

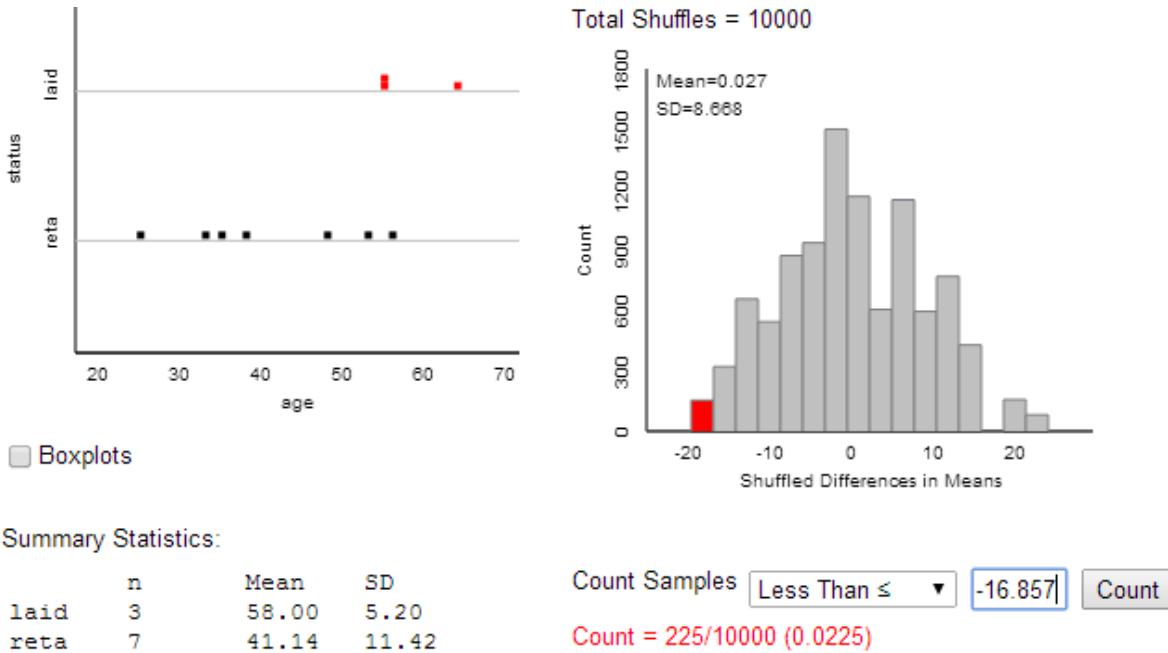
(a) Let  $\mu_{\text{fired}} - \mu_{\text{not fired}}$  represent the difference in the average age of people that would be laid-off, in the long run (by the overall process), and the average age of the people who would be retained. (Don't worry too much at this point about which stage of the firing process this analysis considers. Just keep in mind that we are trying to say something beyond the observed means. We believe there is some underlying difference in means and these data provide an estimate.)

$$H_0: \mu_{\text{fired}} - \mu_{\text{not fired}} = 0 \text{ (no overall difference in the average ages of those getting fired and not)}$$

$$H_a: \mu_{\text{fired}} - \mu_{\text{not fired}} > 0 \text{ (those getting fired will tend to have higher ages than those not)}$$

(b)  $\bar{x}_1 - \bar{x}_2 = 58.00 - 41.14 = 16.86 \text{ years}$

(c) Using the Comparing Groups (Quantitative) applet, the empirical p-value (remembering to match the direction of subtraction, which may vary depending on how you pasted the data in), the output below shows an empirical p-value of 0.0225. Because  $0.0225 > 0.01$ , we would not reject the null hypothesis at the 0.01 level and conclude that this firing process was not more likely to fire individuals with larger ages.



Note: You could also carry out this simulation in R

```
> I = 10000; diff = 0
> for (i in 1:I){
  rerandom = sample(age)
  diff[i] = mean(rerandom[1:3])-mean(rerandom[4:10])
}
```

(d) A two-sample *t*-test in Minitab or R:

```
Two-Sample T-Test and CI: age, fired?
Two-sample T for age
  data: age by fired.
  fired?   N    Mean   StDev   SE Mean
  fired     3    58.00    5.20      3.0
  not fired 7    41.1     11.4      4.3

  Difference = mu (fired) - mu (not fired)
  Estimate for difference: 16.86
  95% lower bound for difference: 6.90
  T-Test of difference = 0 (vs >): T-Value = 3.21  P-Value = 0.007  DF
```

Now  $p\text{-value} < 0.01$ , and we would reject the null hypothesis at the 0.01 level and conclude that this firing process was more likely to fire individuals with larger ages.

(e) The two-sample *t*-test yields a much smaller *p*-value, about 0.006, which implies much stronger evidence of an underlying age difference between the two populations (fired and not fired). However, because of the small sample sizes (especially with the unbalanced groups), we would have major concerns about using the *t*-procedures for this study. The empirical *p*-value from the simulated randomization test is more trustworthy. In fact, in this study, it's probably more work than we need to do, because with only 120 possible combinations it is not unreasonable to calculate the exact *p*-value.

**Example 4.2: Speed Limit Changes**

Try these questions yourself before you use the solutions following to check your answers.

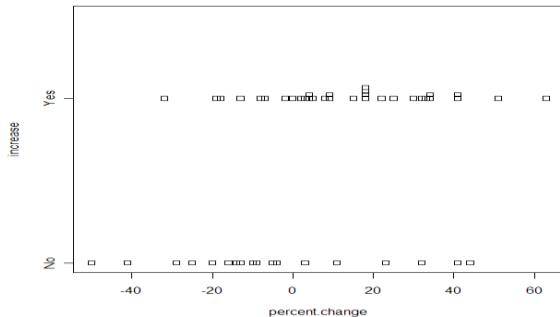
In 1995, the National Highway System Designation Act abolished the federal mandate of 55 miles per hour maximum speed limit and allowed states to establish their own limits. Of the 50 states (plus District of Columbia), 32 increased their speed limits in 1996. The data in [TrafficFatalities.txt](#) shows the percentage change in interstate highway traffic fatalities from 1995 to 1996 and whether or not the state increased their speed limit. (Data from the National Highway Traffic Safety Administration as reported in Ramsey and Schafer, 2002.)

- (a) Identify the observational units and response variable of interest. Is this a randomized experiment or an observational study?
  
  
  
  
  
- (b) Produce numerical and graphical summaries of these data and describe how the two groups compare.
  
  
  
  
  
- (c) Are the technical conditions for a two-sample  $t$ -test met for this study? Explain.
  
  
  
  
  
- (d) Carry out a two-sample  $t$ -test to determine whether the average percentage change in interstate highway traffic fatalities is significantly higher in states that increased their speed limit. If you find a significant difference, estimate its magnitude with a confidence interval.
  
  
  
  
  
- (e) Discuss what the p-value in (d) measures.

## Analysis

(a) The observational units are the 50 states (and the District of Columbia). The response variable of interest is the percentage change in traffic fatalities from 1995 to 1996 (quantitative). This is an observational study because the researchers did not randomly assign which states would increase their speed limits.

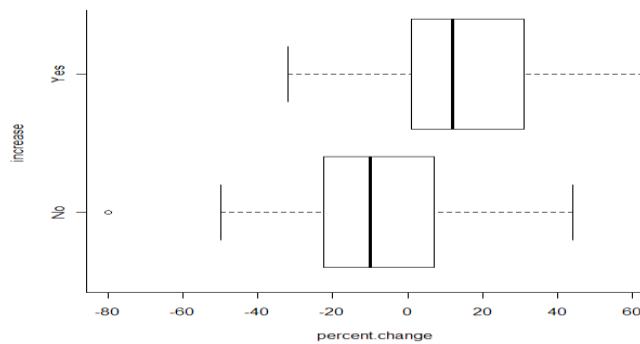
(b) The following graphical display is dotplots of the percentage change in traffic fatalities for each state (and D.C.) in the two groups on the same scale:



Because the distributions are reasonably symmetric, it makes sense to report the means and standard deviations as the numerical summaries:

$$\begin{array}{lll} \text{No increase} & \bar{x}_{\text{no}} = -8.53\% & s_{\text{no}} = 31\% \\ \text{Increase} & \bar{x}_{\text{yes}} = 13.69\% & s_{\text{yes}} = 22\% \end{array}$$

These results indicate that there is a tendency for the percentage change in traffic fatalities to be higher in those states that increase their speed limits. This tendency is also seen in stacked boxplots:



The boxplots also reveal an outlier, the District of Columbia, which did not change its speed limit and had an unusually high decrease in the percentage change of accidents.

These summaries also reveal that the two sample distributions are reasonably similar in shape and variability.

(c) In considering the technical conditions, we see that the sample sizes (19 and 32) are reasonably large. Coupled with the normal shaped sample distributions, the normality/large sample size conditions appears to be satisfied for us to use the *t*-distribution.

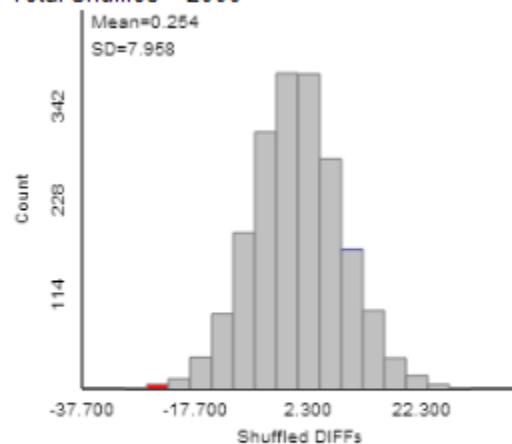
(d) Let  $\mu_{\text{no}} - \mu_{\text{yes}}$  represent the true “effect” of increasing the speed limit on the traffic fatality rate (states that didn’t change speed limit – states that did change speed limit)

$H_0: \mu_{\text{no}} - \mu_{\text{yes}} = 0$  there is no true effect from increasing the speed limit

$H_a: \mu_{\text{no}} - \mu_{\text{yes}} < 0$  increasing the speed limit leads to an increase in traffic fatalities (higher average percentage change with increase in speed limit)

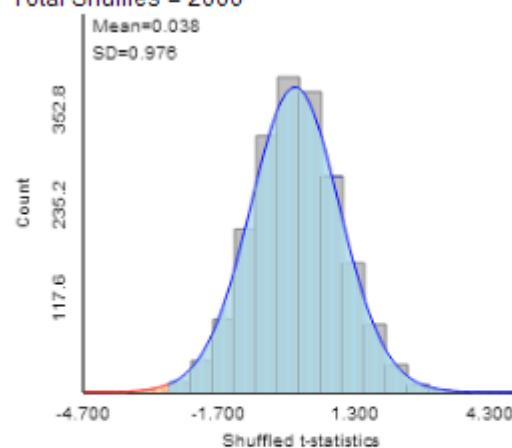
We can apply a randomization test that would look at what would happen if these groups were mixed up with no difference between the “no” group and the “yes” group.

Total Shuffles = 2000



Count Samples [Less Than] -22.214 Count  
Count = 6/2000 (0.0030)

Total Shuffles = 2000



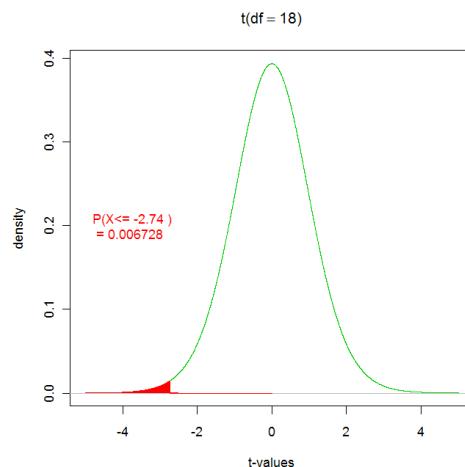
Count Samples [Less Than] -2.76 Count  
Count = 4/2000 (0.0020)  
 Overlay t distribution  
theory-based p-value=0.0050, df = 28.35

We can also approximate this randomization distribution with the two-sample  $t$ -procedure.

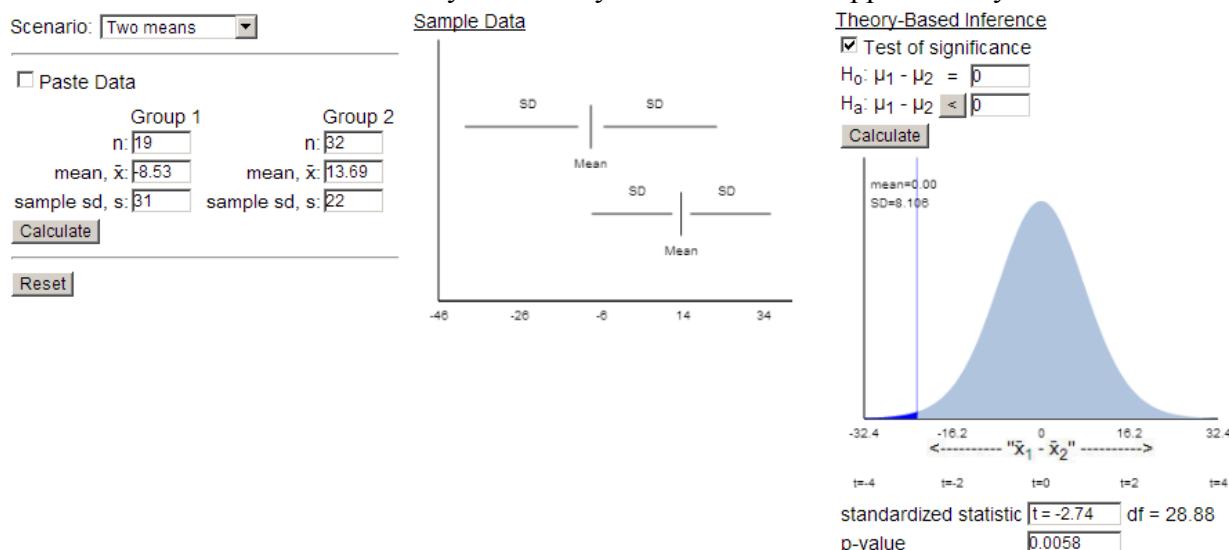
In this case, the (unpooled) test statistic will be  $t = \frac{-8.53 - 13.69}{\sqrt{\frac{31^2}{19} + \frac{22^2}{32}}} = -2.74$

If we approximate the degrees of freedom by  $\min(19 - 1, 32 - 1) = 18$ , then we find the one-sided p-value in R or Minitab to be:

```
> iscamtprob(-2.74, 18, "below")
probability: 0.006728
```



These calculations are confirmed by the Theory-Based Inference applet and by R:



#### Welch Two Sample t-test

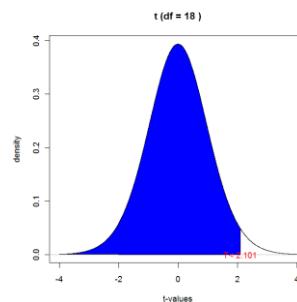
```
data: percent.change by increase
t = -2.7635, df = 28.353, p-value = 0.004968
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -8.545338
sample estimates:
mean in group No mean in group Yes
-8.526316      13.687500
```

Note: R and the Theory-Based Inference applet use a more exact method for determining the degrees of freedom. Our “by hand” method (also used in the applet) is conservative in that the p-value found will be larger than the actual p-value as seen here.

Such a small p-value ( $0.005 < 0.01$ ) reveals that we would observe such a large difference in group means by random assignment alone if there was no treatment effect only about 5 times in 1000, convincing us that the observed difference in the group means is larger than what we would expect just from random assignment. We have strong evidence that something other than “random chance” led to this difference. However, we cannot attribute the difference solely to the speed limit change because this was not actually a randomized experiment. As the states self-selected, there could be confounding variables that help to explain the larger increase in fatality rates in states that increased their speed limit.

Because we rejected the null hypothesis, we are also interested in examining a confidence interval to estimate the size of the treatment effect. We first approximate the  $t^*$  critical value for say 95% confidence, again using  $\min(19 - 1, 32 - 1) = 18$  as the degrees of freedom.

```
> iscaminv(.975, 18, "below")
the observation with 0.975 probability below is
2.101
```



Then the 95% confidence interval can be calculated,

$$-8.53 - 13.69 \pm 2.10092 \sqrt{\frac{31^2}{19} + \frac{21^2}{32}} = -22.2 \pm 16.85$$

We are 95% confident that the true “treatment effect” is in this interval or that the mean percentage increase in traffic fatality rates is between 5.4 percentage points to 39.1 percentage points higher in states that increase their speed limit compared to states that do not increase their speed limit (continuing to be careful not to state this as a cause and effect relationship).

Before we complete this analysis, it is worthwhile to investigate the amount of influence that the outlier (the District of Columbia) has on the results, especially because D.C. does have different characteristics from the states in general. The updated R output (two-sided p-value) is below:

#### Welch Two Sample t-test

```
data: percent.change by increase
t = -2.5079, df = 29.687, p-value = 0.01785
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-33.105399 -3.380712
sample estimates:
mean in group No mean in group Yes
-4.555556 13.687500
```

As we might have guessed, the mean increase in fatalities for the “No” group has increased so that the difference in the group means is less extreme. This leads to a less extreme test statistic and a larger p-value (one-sided p-value =  $0.01785/2 = 0.0089$ ) so somewhat weaker evidence against the null hypothesis in favor of the one-sided alternative hypothesis.

(e) The other technical condition is that we have independent random samples or random assignment to groups. We do not have either in this study, because we are examining the population of all states (and D.C.), and the states self-selected whether they changed their speed limit. Thus, any p-value we calculate is in a sense hypothetical because we have all the states here, we might ask the question: would the two groups look this different if whether or not they increased their speed limit had been assigned at random?

So the above p-value measures how often we would see a difference in group means at least this large based on random assignment to the two groups if there were no true treatment effect. Even though this p-value is hypothetical, we still have some sense that the difference observed between the groups is larger than we would expect to see “by chance” even in a situation like this where it is not feasible to carry out a true randomized experiment. This gives some information that can be used in policy decisions but we must be careful not to overstate the attribution to the speed limit change.

**Example 4.3: Distracted Driving?**

Try these questions yourself before you use the solutions following to check your answers.

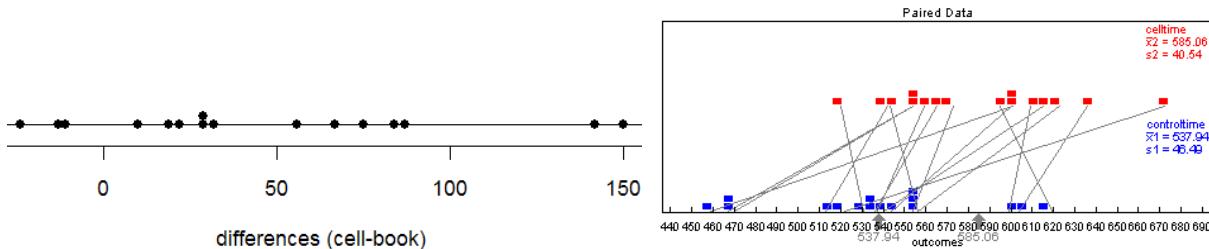
Recall the Distracted Driving study from Example 2.2. The reaction times (in milliseconds) for 16 students appear below and in the file [driving.txt](#).

Subject	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Cell	636	623	615	672	601	600	542	554	543	520	609	559	595	565	573	554
Control	604	556	540	522	459	544	513	470	556	531	599	537	619	536	554	467

- (a) Analyze the *differences* in reaction times (cell phone minus control) for these subjects. Include numerical and graphical summaries of the distribution of differences. Comment on what this descriptive analysis reveals about whether talking on a cell phone tends to produce slower reaction times.
- (b) State the appropriate null and alternative hypotheses to be tested, in order to investigate the research question of whether talking on a cell phone tends to produce slower reaction times.
- (c) Conduct a simulation analysis of a randomization test for testing these hypotheses. Report the empirical p-value. Summarize the conclusion that you would draw from this analysis.
- (d) Comment on whether the conditions for applying a paired *t*-test and *t*-interval are satisfied for these data.
- (e) Conduct a paired *t*-test of these hypotheses. Report the value of the test statistic and the p-value. Indicate your test decision at the 0.05 and 0.01 significance levels, and summarize your conclusion.
- (f) Produce and interpret a 95% *t*-confidence interval for the population mean difference. Also produce and interpret a 95% prediction interval. Comment on how these two intervals compare.
- (g) Compare these analyses to that in Example 2.2. Which analysis would you recommend and why?

## Analysis

### (a) Analyzing the differences

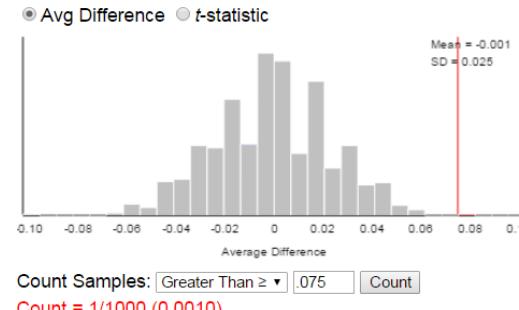


The sample mean difference in reaction times (cell minus control) is  $\bar{x}_d = 47.125$  milliseconds, with a standard deviation of  $s_d = 51.331$  milliseconds. The dotplot reveals that most of the differences are positive, suggesting that subjects talking on a cell phone tend to take longer to react than subjects listening to a book-on-tape.

(b) The null hypothesis says that the mean reaction time is the same among cell phone users as among book-on-tape listeners ( $H_0: \mu_{\text{cell}} - \mu_{\text{control}} = 0$ ). The alternative says that the mean reaction time is larger among cell phone users than among book-on-tape listeners ( $H_a: \mu_{\text{cell}} - \mu_{\text{control}} > 0$ ).

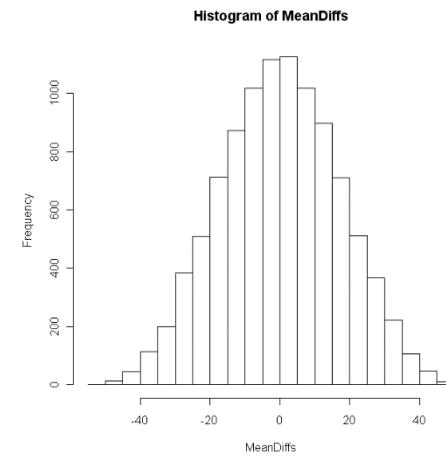
(c) We can carry out the simulation easily with the applet or with R.

Copying and pasting the data into the Matched Pairs applet with 1,000 repetitions we get the results shown here.



Using R to carry out the simulation instead:

```
MeanDiffs=0
for (i in 1:10000) {
  multiplier=sample(c(-1,1), 16,
+ replace=TRUE)
  RandomizedData=differences*multiplier
  MeanDiffs[i]=mean(RandomizedData)
}
```



**From R:** The empirical p-value is the proportion of these 10,000 repetitions in which the mean difference is 47.125 or more, because 47.125 is the value of the sample mean difference from the actual experimental data. None of the 10,000 repetitions produced such a large mean difference, so the empirical p-value is 0. The simulation therefore shows that we would almost never get a result as extreme as the actual experiment did, if there were really no difference between reactions to cell phone

vs. book-on-tape, so we have extremely strong evidence that the cell phone really does increase reaction times.

(d) Because the sample size (16) is fairly small, the *t*-procedures are valid only if the population of differences follows a normal distribution. The dotplot of differences from these 16 subjects looks roughly symmetric, so the *t*-procedures are probably valid to apply here.

(e) The test statistic is  $\frac{\bar{x}_d - 0}{s_d / \sqrt{n_d}} = \frac{47.125}{51.331 / \sqrt{16}} \approx 3.67$ . The p-value is the probability that a *t*-

distribution with 15 degrees of freedom is 3.67 or larger; R reveals this p-value to be 0.001137. This p-value is very small, so we would reject the null hypothesis at the 0.05 and 0.01 significance levels. The experimental data provide very strong evidence that talking on a cell phone does cause an increase in mean reaction time, as compared to listening to a book-on-tape. The cause/effect conclusion is justified because this is a randomized experiment with a very small p-value.

Note: This evidence is stronger than what we saw with the sign test in Example 2.2.

(f) A 95% confidence interval for the population mean difference  $\mu_d$  is:  $\bar{x}_d \pm t^* \times s_d / \sqrt{n_d}$ , which is  $47.125 \pm 2.131(51.331) / \sqrt{16}$ , which is  $47.125 \pm 27.347$ , which is (19.778, 74.472). We can be 95% confident that the mean reaction time while talking on a cell phone is between roughly 20 and 75 milliseconds longer than when listening to a book-on-tape.

A 95% *prediction* interval for the difference in reaction times for a particular subject is:

$\bar{x}_d \pm t^* s_d \sqrt{1 + 1/n_d}$ , which is  $47.125 \pm 2.131(51.331) \sqrt{1 + 1/16}$ , which is  $47.125 \pm 112.753$ , which is (-64.628, 159.878). We can be 95% confident that an individual subject will react anywhere from 65 milliseconds more quickly to 160 milliseconds more slowly talking on a cell phone as compared to listening to a book-on-tape.

(g) An advantage of the sign test from Chapter 2 is we don't need to rely on large sample sizes or normality of the differences for the procedure to be valid. However, by ignoring some information in the data, we would expect this procedure to have lower power and we lose information about the size of the difference. The conditions for the *t*-procedure seem to be satisfactorily met here.

## CHAPTER 4 SUMMARY

In this chapter, you have again focused on comparing the outcomes for different groups, but for quantitative data. The tools for descriptive statistics are the same as in chapter 2 (e.g., means, medians, standard deviations, interquartile ranges, dotplots, histograms, and boxplots).

After exploring and comparing the groups descriptively, it may be appropriate to ask whether the differences observed could have arisen “by chance alone.” In other words, is the observed difference large enough to convince us that it arose from a genuine difference in the groups instead of from the randomness inherent in designing the experiment or in selecting the samples? Analogous to comparing proportions in Chapter 3, we used simulation to approximate how often we would obtain a difference in means at least as extreme as observed just by chance (random sampling or random assignment). Then we learned of a “large sample” approach we can use in either case to approximate the null distribution of the standardized statistic, yielding the two-sample  $t$ -test. Keep in mind that the logic of statistical significance hasn’t changed – we assume there is no effect or no population difference and determine how often we would get a sample difference result at least as extreme as what was observed. You saw that this chance is strongly affected by the amount of variability within the groups. It is very important when reporting on your statistical study that you also mention the sample sizes. This helps you evaluate the power and practical significance of the study results. You were also reminded that no matter how small a p-value is, we cannot draw cause-and-effect conclusions unless the data came from a properly designed randomized experiment.

You will often have the choice between a randomization-based test or the  $t$ -procedures. The advantage to the randomization procedures is they allow us to work with many types of statistics (such as median and trimmed mean) where we do not have theoretical results on the large sample behavior of the randomization distribution. The advantage of the  $t$ -procedures is the ease in calculating confidence intervals. It is always important to consider the technical conditions of an inference procedure before applying the procedure (e.g., make sure you really have two independent samples). If data are paired, then you should consider one-sample procedures on the differences. Such pairing is often very useful in increasing the efficiency of the study. These  $t$ -procedures are quite robust in calculating p-values and confidence intervals.

Always keep in mind the importance of considering whether there is random sampling and/or random assignment in the study design and how that impacts the scope of conclusions that you can draw from your study.

### SUMMARY OF WHAT YOU HAVE LEARNED IN THIS CHAPTER

- Stating hypotheses in terms of comparing two population or underlying treatment means
- The reasoning process behind a randomization test
  - Calculating and interpreting p-values
- Two-sample  $t$ -procedures for comparing two groups
  - When appropriate to apply, why helpful
  - How to interpret results
  - Alternative approaches if not valid (e.g., data transformations)
- Using technology to calculate two-sample  $t$ -procedures

- Factors that affect the significance of the comparison (e.g., sample size, size of difference, and within-group variability)
- Methods for reducing within-group variability
- Matched-pairs designs and analyses

## TECHNOLOGY SUMMARY

- Simulating a randomization test for different statistics
- Two-sample  $t$ -procedures
- Paired  $t$ -procedures

## Quick Reference to ISCAM Workspace Functions and other R Commands

Procedure Desired	Function Name (options)
Stacked Dotplot	iscamdotplot (response, explanatory, names)
Stack data	stack(data[, c("var1", "var2")])
Parallel Dotplots (stacked data)	iscamdotplot (response, explanatory)
Parallel Histograms (stacked data)	Load Lattice package, then histogram (~response   explanatory, layout=c(1,2))
Parallel Boxplots (stacked data)	boxplot(response~explanatory, horizontal=TRUE)
Numerical Summaries	iscamsummary(response, explanatory)
two-sample $t$ -procedures	t.test(response~explanatory, alt, var.equal=FALSE)
two-sample $t$ -procedures (summary data)	iscamtowosamplet(x1, s1, n1, x2, s2, n2, hypothesized, alternative, conf.level)
Paired $t$ -test	t.test(v1, v2, alternative, conf.level, paired=TRUE)

## Quick Reference to Minitab Commands

Procedure Desired	Menu
Stacked dotplots (stacked data)	Graph > Dotplot, One Y, With Groups
Stacked dotplots (unstacked)	Graph > Dotplot, Multiple Y's, Simple
Stack data	Data > Stack > Columns Or MTB> stack c1 c2 c3; SUBC> subs c4.
Parallel Histograms (unstacked)	Graph > Histogram, Simple, Multiple Graphs
Numerical Summaries	Stat > Basic Statistics > Display Descriptive Statistics with By Variable
two-sample $t$ -procedures (stacked data)	Stat > Basic Statistics > 2-Sample t, Samples in one column
two-sample $t$ -procedures (unstacked data)	Stat > Basic Statistics > 2-Sample t, Samples in different columns
two-sample $t$ -procedures (summary data)	Stat > Basic Statistics > 2-Sample t, Summarized data
Paired $t$ test	Stat > Basic Statistics > Paired t

## Choice of Procedures for Comparing Two Means

Parameter	Difference in population means or treatment means ( $\mu_1 - \mu_2$ )	Difference in population medians or treatment medians	Mean difference
<b>Study design</b>	Randomized experiment or independent random samples	Randomized experiment or independent random samples	Matched paired design
<b>Null Hypothesis</b>	$H_0: \mu_1 - \mu_2 = 0$	$H_0:$ population medians equal	$H_0: \mu_{diff} = 0$
<b>Simulation</b>	Randomly assign response values between groups (see Comparing Groups (Quantitative) applet)		Flip a coin to see whether to flip the order of the observations
<b>Exact p-value</b>	All possible random assignments		
<b>Valid to use <math>t</math>-procedures if</b>	Both sample sizes at least 30 or both population distributions normal	N/A	At least 30 differences or differences are normal
<b>Test Statistic</b>	$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	N/A	$t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n_d}}$
<b>Confidence interval</b>	$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$		$\bar{x} \pm t^* s / \sqrt{n}$
<b>Prediction interval</b>			$\bar{x} \pm t^* s / \sqrt{1 + 1/n}$
<b>R Commands</b>	<pre>iscamtwosampel • mean1, stddev1, n1, mean2, stddev2, n2 • Optional: hypothesized difference and alternative ("less", "greater", or "two.sided") • Optional: conf.level</pre> <p><code>t.test(resp~exp, alt, var.equal=FALSE)</code></p>		<code>t.test(list1, list2, alt, conf.level, paired=TRUE)</code>
<b>Minitab</b>	Stat > Basic Statistics > 2-sample t		Stat > Basic Statistics > Paired t
<b>TBI applet</b>	Two means		

**Note:** You can also consider transforming the data before applying normal-based methods

## CHAPTER 5: COMPARING SEVERAL POPULATIONS, EXPLORING RELATIONSHIPS

The idea of comparing two groups has been a recurring theme throughout this course. In the previous chapters, you have been limited to exploring two groups at a time. You saw that often the same analysis techniques apply whether the data have been collected as independent random samples or from a randomized experiment, although this data collection distinction strongly influences the scope of conclusions that you can draw from the study. You will see a similar pattern in this chapter as you extend your analyses to exploring two or more groups. In particular, you will study a procedure for comparing a categorical response variable across several groups and a procedure for comparing a quantitative response variable across several groups. You will also study the important notion of association between variables, first with categorical variables and then for studies in which both variables are quantitative. In this latter case, you will also learn a new set of numerical and graphical summaries for describing these relationships.

### Section 1: Two Categorical Variables

- Investigation 5.1: Dr. Spock's trial – Chi-square test for homogeneity of proportions
- Investigation 5.2: Nightlights and near-sightedness (cont.) – Chi-square test for association
- Technology Exploration: Randomization test for chi-square statistic
- Investigation 5.3: Newspaper credibility decline – Comparing distributions

### Section 2: Comparing Several Population Means

- Investigation 5.4: Disability discrimination – Reasoning of ANOVA
- Applet Exploration: Randomization test for ANOVA
- Investigation 5.5: Restaurant spending and music – ANOVA practice
- Applet Exploration: Exploring ANOVA

### Section 3: Two Quantitative Variables

- Investigation 5.6: Cat jumping – Scatterplots
- Investigation 5.7: Drive for show, putt for dough – Correlation coefficients
- Applet Exploration: Correlation guessing game
- Investigation 5.8: Height and foot size – Least squares regression
- Applet Exploration: Behavior of regression lines – Resistance
- Excel Exploration: Minimization criteria
- Investigation 5.9: Money-making movies – Application

### Section 4: Inference for Regression

- Investigation 5.10: Running out of time – Inference for regression (sampling)
- Investigation 5.11: Running out of time (cont.) – Inference for regression (shuffling)
- Investigation 5.12: Boys' heights – Regression model
- Investigation 5.13: Cat jumping (cont.) – Confidence intervals for regression
- Investigation 5.14: Housing prices – Transformations
- Technology Exploration: The regression effect

- Example 5.1: Internet Use by Region
- Example 5.2: Lifetimes of Notables
- Example 5.3: Physical Education Class Performance
- Example 5.4: Comparing Popular Diets

## SECTION 1: TWO CATEGORICAL VARIABLES

In Chapter 3 you learned inference procedures for assessing whether two population proportions are equal and whether two categorical variables are independent. Those methods were limited to *binary* variables. In this section, you will expand on your earlier techniques to allow for more than two categories in each categorical variable.

### Investigation 5.1: Dr. Spock's Trial

The well-known pediatrician and child development author Dr. Benjamin Spock was also an anti-Vietnam War activist. In 1968 he was put on trial and convicted on charges of conspiracy to violate the Selective Service Act (encouraging young men to avoid the draft). The case was tried by Judge Ford in Boston's Federal court house. A peculiar aspect of this case was that his jury contained no women. A lawyer writing about the case that same year in the Chicago Law Review said, "Of all defendants at such trials, Dr. Spock, who had given wise and welcome advice on child-bearing to millions of mothers, would have liked women on his jury" (Ziesel, 1969). The opinion polls also showed that women were generally more opposed to the Vietnam War than men.

In the Boston District Court, jurors are selected in three stages. The Clerk of the Court is supposed to select 300 names at random from the City Directory and put a slip with each of these names into a box. The City Directory is renewed annually by a census of households visited by the police, and it lists all adult individuals in the Boston area. In Dr. Spock's trial, this sample included only 102 women, even though 53% of the eligible jurors in the district were female. At the next stage, the judge selects 30 or more names from those in the box which will constitute the "venire." Judge Ford chose 100 potential jurors out of these 300 people. His choices included only 9 women. Finally, 12 actual jurors are selected after interrogation by both the prosecutor and the defense counsel. Only one potential female juror came before the court and she was dismissed by the prosecution.

In filing his appeal, Spock's lawyers argued that Judge Ford had a history of venires in which women were systematically underrepresented. They compared the gender breakdown of this judge's venires with the venires of six other judges in the same Boston court from a recent sample of court cases. Records revealed the following data:

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Total
<b>Women on jury list</b>	119	197	118	77	30	149	86	776
<b>Men on jury list</b>	235	533	287	149	81	403	511	2199
<b>Total</b>	354	730	405	226	111	552	597	2975

#### Descriptive Statistics

- (a) Calculate the proportion of women on the jury list for each judge. Also create a [segmented bar graph](#) to compare these distributions. How do the judges compare?

- (b) Let  $\pi_i$  represent the probability of judge  $i$  selecting a female for the jury list. State a null and an alternative hypothesis for testing whether these data provide reason to doubt that the proportion of women on jury lists is the same for all seven judges.

*Note:* Your null hypothesis states only that the probabilities are equal; you are not specifying a particular value for this common probability. The alternative hypothesis can state only that at least one probability differs from the rest.

If the null hypothesis is true, then the long-run probability of a woman on the jury panel equals the same value for all seven judges.

### Null Model

- (c) Suggest an estimate for this common (across all judges) probability that a juror is female.

- (d) How many jurors were on Judge 1's list? If you suppose the long-run proportion of women in his juries was 0.261, how many of these jurors would you expect to be female? How many would you expect to be male? (*Recall:* An “expected value” does not need to be an integer.)

Female:

Male:

- (e) How many of the jurors on Judge 2's list would you expect to be female if his long-run proportion was also 0.261? How many would you expect to be male?

Female:

Male:

- (f) Enter your expected counts below the observed counts in the following table.

		Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7
Women on jury list	Observed	119	197	118	77	30	149	86
	Expected			105.71	58.99	28.97	144.07	155.82
Men on jury list	Observed	235	533	287	149	81	403	511
	Expected			299.30	167.01	82.03	407.93	441.18
	Total	354	730	405	226	111	552	597

(g) Are the observed counts equal to the expected counts in each cell of the table? Is it possible that the long-run probability of a female jury panel member is the same for each judge, and the differences between the observed counts and the expected counts found in the table are due to random chance alone?

### Test Statistic

(h) Suggest a statistic for measuring the overall deviation between the observed counts and the expected counts. [Hint: Write your statistic as a formula or rule for obtaining *one* number that takes into account information in the sample relevant to comparing all seven groups.]

A common test statistic used to compare the observed and expected counts in a two-way table is the [Chi-squared test statistic](#):

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

where  $r$  = number of rows and  $c$  = number of columns. This test statistic looks at the discrepancy between the observed and expected counts for each cell, squares the difference to ensure a positive contribution, “standardizes” by dividing by the expected count, and then sums across all cells.

(i) Calculate this test statistic for the above two-way table. [Hint: Fill in the terms for the first two judges and then sum all of values together.]

$$\begin{aligned}\chi^2 = & \quad + \quad + \frac{(118 - 105.71)^2}{105.71} + \frac{(77 - 58.99)^2}{58.99} + \frac{(30 - 28.97)^2}{28.97} + \frac{(149 - 144.07)^2}{144.07} + \frac{(86 - 155.82)^2}{155.82} \\ & + \quad + \quad + \frac{(287 - 299.30)^2}{299.30} + \frac{(149 - 167.01)^2}{167.01} + \frac{(81 - 82.03)^2}{82.03} + \frac{(403 - 407.93)^2}{407.93} + \frac{(511 - 441.18)^2}{441.18}\end{aligned}$$

$$\begin{aligned}\chi^2 = & \quad + \quad + 1.45 + 5.53 + 0.04 + 0.17 + 31.22 \\ & + \quad + \quad + 0.51 + 1.95 + 0.01 + 0.06 + 11.02 =\end{aligned}$$

(j) What types of test statistic values (large, small, positive, negative) constitute evidence against the null hypothesis of equal long-run probabilities? Explain.

### Simulation

In order to approximate the p-value of this test, we need to examine how the test statistic varies under the null hypothesis of equal probabilities. We will once again explore this null distribution first by simulating a large number of random samples where the probability of a juror being female is the same for each judge (under the null hypothesis).

(k) Outline the steps you would use to generate random data for each judge under the null hypothesis that the probability of a juror being female is the same for each judge.

(l) Use a premade script or macro to carry out the simulation:

- **In R:** Run the [Spock.R](#) script to generate 1000 chi-square values.

(m) Describe the shape, center, and variability of the simulated null distribution.

(n) Based on your simulation, determine the proportion of simulated test statistic values that are as large or larger than the test statistic value you computed in (i). Does this empirical p-value provide convincing evidence that the difference between observed and expected counts that you observed is larger than we would expect by chance? What do you conclude about whether the seven judges all had the same long-run probability of a juror being female? Explain.

### Mathematical Model

Rather than rely on simulation to produce (approximate) p-values for this test, we can use a probability model to approximate the sampling distribution of the chi-square test statistic. We just need to find the right one.

(o) Use technology to determine whether a normal probability distribution appears to adequately predict the behavior of the simulated null distribution of this test statistic:

- **In R:** To create a normal probability plot: `> qqnorm(chisqvalues)`

What is your assessment?

(p) Use technology to whether a “chi-square” probability model appears to adequately predict the behavior of the simulated null distribution of this test statistic:

Create a probability plot for the chi-square distribution with 6 degrees of freedom.

- **In R:** Create a qqplot with data randomly generated from a chi-square distribution  
`> qqplot(chisqsum, qchisq(ppoints(1000), df = 6))`

(q) What is your assessment of the fit of this model to the simulated test statistic values?

The [Chi-Square distribution](#) is skewed to the right and provides a reasonable model to the above test statistic for large sample sizes. We will consider the chi-square distribution model appropriate if all of the expected counts are at least 1 and if at least 80% of the expected counts are at least 5.

When comparing several population proportions, the chi-square [degrees of freedom](#) are equal to the number of explanatory variable categories minus 1,  $c - 1$ . This makes sense because once we specify the number of observations in  $c - 1$  of the categories, the last category is forced to assume the value that allows the observed counts to sum to the sample size.

For large sample sizes, we will use the chi-square distribution to approximate the p-value.

### Technology Detour – Chi-square Probabilities

**In R:** The `iscamchisqprob` function takes the following inputs

- $xval$  = the observed value of the test statistic
- $df$  = degrees of freedom for the chi-square distribution

(r) How does this p-value compare to the empirical p-value you determined in (n)?

**Discussion:** If the null hypothesis is rejected, the conclusion we draw is that at least one of the population proportions differs from the rest, but we don’t have much information about which one. It could be that one explanatory variable group is behaving much differently than the rest or they could all be different. One way to gain more information about the nature of the differences between the  $\pi_i$  values is to compare the components of the chi-square statistic sum.

(s) Return to the sum you calculated in (i). Which cell comparison(s) provide the largest (standardized) discrepancy between the observed counts and the expected counts?

(t) For the cells identified in (s), which is larger, the observed counts or the expected counts? Explain the implications of this comparison.

- (u) Which judge do you believe tried Dr. Spock's case? Explain.

### Study Conclusions

One judge clearly stood out compared to the others in these sample data. If we consider these results to be representative of the overall jury selection process, the very small p-value indicates that if in fact the judges' selections of jurors were independent random processes with the same probability of selecting a woman, then it would be almost impossible to observe sample proportions differing by this much by chance alone. Thus, the sample data provide strong evidence that the long-run probability of a juror being female is not the same among all seven judges. The largest contributions to the  $\chi^2$  test statistic, by far, come from judge 7, who has many more men than would be expected and many fewer women than would be expected on his jury lists. This was indeed Judge Ford, the judge assigned to Dr. Spock's case. In fact, there are two issues with this judge: The sampling from the city directory led to a far smaller percentage of women (29%) than the city population (53%) across all the judges, and then the proportion of women selected by Judge Ford dipped even lower to around 15% women. (By the way, the Court of Appeals reversed Spock's conviction on other grounds without reaching the jury selection issue.) Although this p-value is hypothetical in nature (there was not a true random mechanism used in generating these outcomes), and we cannot draw a cause-and-effect conclusion because of the observational nature of the data, the p-value does provides a measure of how very surprising these results would be to occur by chance alone.

### Technical Conditions

The chi-square distribution is an approximation to the sampling distribution of the chi-square test statistic when the data arise from independent binomial random variables. This approximation is considered valid as long as the average expected cell count is at least 5 and all of the individual expected cell counts are at least one. Notice we are discussing the *expected* cell counts here, not the observed cell counts. The data also need to have been collected from independent random samples or from a randomized comparative experiment.

The advantage of a Chi-Square Test is that it provides an overall p-value for all the comparisons at once which controls the overall probability of a Type I error. If the p-value is not significant, we usually do not check all of the individual comparisons. If the p-value is significant, then we can do more formal follow-up analyses to see where the difference(s) are arising. If we run many tests on the same data set, we are always concerned about an inflated overall Type I error rate.

## Summary of Chi-Square Test of Homogeneity of Proportions

Numerical and graphical summaries: conditional proportions and segmented bar graphs

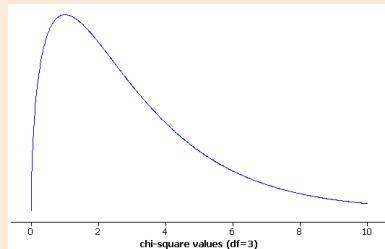
If the data are independent random samples from  $I$  different populations or processes and at least 80% of the expected cell counts are at least 5 (and all are at least one), then the hypotheses

$H_0: \pi_1 = \dots = \pi_I$  where  $\pi_i$  is the probability of success in population  $i$ .

$H_a:$  at least one  $\pi_i$  differs from the rest

can be tested using the chi-square test statistic:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^I \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$



The (upper-tail) p-value is calculated from the chi-square distribution with  $(I - 1)$  degrees of freedom.

Note: The chi-square calculations are not affected by which variable you treat as explanatory and which as response.

### Practice Problem 5.1

One question that you may be asking is why not just use the two-sample  $z$  procedures to compare pairs of proportions? We could do this for two groups at a time, e.g., comparing Judge 1 to Judge 2 and then Judge 3 to Judge 5 and so on.

- (a) How many such two-group comparisons are there among these 7 judges?
- (b) If the level of significance is set to 0.05, what is the probability of a Type I error for any one of these comparisons?
- (c) What about the probability of *at least one* Type I error among these 21 comparisons – will that be larger or smaller than 0.05? Explain.

## Technology Detour – Chi-Square Tests

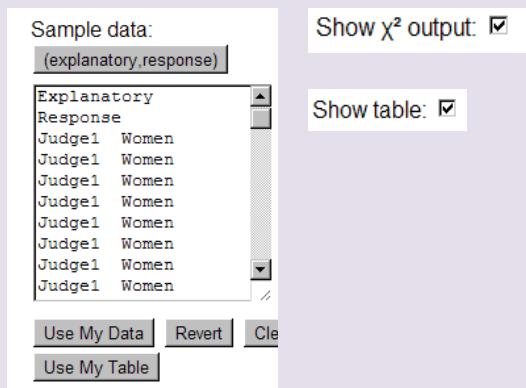
### In R

- If you have the counts for the two-way table, can pass the table using the matrix command:
- ```
> spocktable = matrix(c(119, 235, 197, 533, 118, 287, 77, 149,
  30, 81, 149, 403, 86, 511), ncol=7)
> chisq.test(spocktable)
```
- For the raw data (e.g., in spock), create and pass in the frequency table, e.g.,
`> chisq.test(table(spock))`

### In Analyzing Two-way Tables applet

- Paste either the raw data (press **Use Data**) or the two-way table (press **Use Table**) into the **Sample data** box, using one word category names.
- Check the **Show X<sup>2</sup> Output** box.

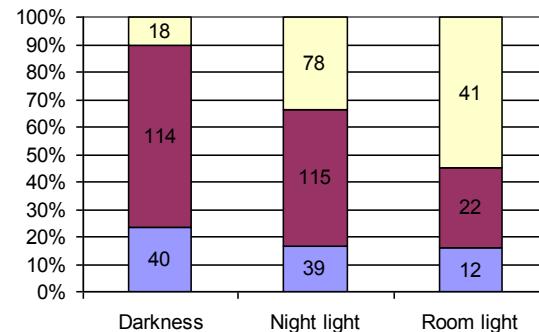
The applet reports the observed  $X^2$  value, df, and cell contributions. (You can expand this window by dragging out the lower right corner. Check the **Show table** box to see the observed counts.)



### Investigation 5.2: Nightlights and Near-sightedness (cont.)

Recall from Investigation 3.2 we examined a simplified version of the study published by Quinn, Shin, Maguire and Stone (1999) that examined the relationship between the type of lighting children were exposed to and subsequent eye refraction. Now we are able to look at 3 categories for each variable. Below is the two way table and segmented bar graph from the study:

|                     | <b>Dark</b> | <b>Night light</b> | <b>Room light</b> | <b>Total</b> |
|---------------------|-------------|--------------------|-------------------|--------------|
| <b>Far-sighted</b>  | 40          | 39                 | 12                | 91           |
| <b>Normal</b>       | 114         | 115                | 22                | 251          |
| <b>Near-sighted</b> | 18          | 78                 | 41                | 137          |
| <b>Total</b>        | 172         | 232                | 75                | 479          |



(a) What proportion of subjects in this study were classified as being far-sighted? As having normal vision? As being near-sighted? [Hint: Looking at the subjects all together now.]

Far-sighted:

Normal:

Near-sighted:

(b) If the children's current eye conditions were not related to the level of lighting they were exposed to, what proportion of children in the darkness condition would be far-sighted? Have normal vision? Be near-sighted?

(c) What would the proportional breakdown of eye conditions look like in the night-light group and in the room-light group if there was no association between eye condition and lighting level?

(d) Use your answers in (b) and (c) to calculate the *expected number* of children in each cell of the table, assuming that there is no relationship between lighting and eye refraction. [Hint: For example, multiply the number of children in the darkness group by the proportion who would have hyperopia in that group if there were no relationship.] Record your answers in the table below next to the observed counts (the diagonal entries have been done for you).

*Expected counts:*

|                     | <b>Darkness</b> | <b>Night light</b> | <b>Room light</b> | <b>Total</b> |
|---------------------|-----------------|--------------------|-------------------|--------------|
| <b>Far-sighted</b>  | 40 ( 32.68 )    | 39 ( )             | 12 ( )            | 91           |
| <b>Normal</b>       | 114 ( )         | 115 ( 121.57 )     | 22 ( )            | 251          |
| <b>Near-sighted</b> | 18 ( )          | 78 ( )             | 41 ( 21.45 )      | 137          |
| <b>Total</b>        | 172             | 232                | 75                | 479          |

If we assume there is no relationship between the explanatory variable and the response variable, we can calculate the *expected count* for each cell by calculating

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{\text{Table total}}$$

This produces the same distribution of conditional proportions in each explanatory variable group. This formula generalizes the approach you used in Investigation 5.1 to tables with more than 2 rows and more than 2 columns.

(e) Are the observed counts equal to the expected counts in each cell of the above table? Is it possible, if there were no relationship between lighting condition and eyesight, that we might have observed differences this large just by random chance?

(f) Use technology to calculate a chi-square test statistic to measure the discrepancy between the observed and expected counts for this table

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

where  $r$  = number of rows and  $c$  = number of columns. (See the previous technology detour to carry out these calculations in R or Minitab.) What are the values of the chi-square statistic, degrees of freedom, and p-value? [You can also verify the expected counts that you computed in part (d).]

*In R: chisq.test(data)\$expected]*

Chi-square statistic:

df:

p-value:

The degrees of freedom for the chi-square test statistic with  $r$  rows and  $c$  columns is

$$df = (r - 1)(c - 1)$$

(g) Minitab and R also report the chi-square contributions for each cell (*In R: chisq.test(data)\$residuals* for the unsquared versions). Which cell(s) contribute the most to the overall chi-square sum? Compare the observed counts to the expected counts for those cells. What do these comparisons reveal about the nature of the relationship between children's eye condition and level of lighting?

### Study Conclusions

The segmented bar graph reveals that for the children in this sample the incidence of near-sightedness increases as the level of lighting increases. Because the observed counts are large, we can apply the chi-square test to these data. The p-value of this chi-square test is essentially zero, which says that if there were no association between eye condition and lighting in the population, then it's virtually impossible for chance alone to produce a table in which the conditional distributions would differ by as much as they did in the actual study. Thus, the sample data provide overwhelming evidence that there is indeed an association between eye condition and lighting in the population of children. A closer analysis of the table and the chi-square calculation reveals that there are many fewer children with near-sightedness than would be expected in the "darkness" group and many more children with near-sightedness than would be expected in the "room light" group. But remember the main lesson of this study from Chapter 3 – we cannot draw a cause-and-effect conclusion between lighting and eye condition because this is an observational study. Several confounding variables could explain the observed association. For example, perhaps near-sighted children tend to have near-sighted parents who prefer to leave a light on because of their own vision difficulties, while also passing this genetic predisposition on to their children. We also have to be careful in generalizing from this sample to a larger population because the children were making voluntary visits to an eye doctor and were not selected at random from a larger population.

### Technical Conditions

The sample size condition to be checked by the expected cell counts for this chi-square procedure is the same as in Investigation 5.1: At least 80% of the expected cell counts are at least 5 and all of the individual expected cell counts are at least one. To generalize to the larger population, we also need the data to be a simple random sample from the population or process of interest, with the observational units cross-classified by the two categorical variables.

### Practice Problem 5.2

The *National Vital Statistics Reports* provided data on gestation period for babies born in 2002. The following table classifies the births by the mother's race and by the duration of the pregnancy:

|                                  | <b>White (non-Hispanic)</b> | <b>Black (non-Hispanic)</b> | <b>Hispanic</b> |
|----------------------------------|-----------------------------|-----------------------------|-----------------|
| <b>Pre-term (under 37 weeks)</b> | 251,132                     | 101,423                     | 99,510          |
| <b>Full term (37 - 42 weeks)</b> | 1,885,189                   | 435,923                     | 692,314         |
| <b>Post-term (over 42 weeks)</b> | 149,898                     | 36,896                      | 64,997          |

- (a) Consider these observations as a random sample from the birth process in the U.S. and conduct a chi-square test of whether these data suggest an association between race and length of gestation period. Report the hypotheses, validity of technical conditions, sketch of sampling distribution, test statistic, and p-value. [Provide the details of your calculations and/or relevant computer output.] Summarize your conclusion.
- (b) Which 2-3 of the nine cells in the table contribute the most to the calculation of the  $\chi^2$  test statistic? Is the observed count lower or higher than the expected count in those cells? Summarize what this reveals about the association between race and length of gestation period.

### Technology Exploration: Randomization Test for Chi-square Statistic

What if our student involved random assignment rather than random sampling? How could we carry out a simulation analysis?

Suppose we wanted to analyze the Night Light study as a pseudo experiment and wanted to explore the treatment effects of the lighting variable.

(h) Outline the design of a simulation that will explore whether the association observed in our sample data could plausibly have arisen from this random assignment process alone, using the chi-square test statistic. [Hint: Consider the data file with two columns, one for the response variable and one for the explanatory variable.]

(i) Use technology to carry out this simulation.

- **In R:** Create a function to store the chisq values (first initialize the vector), such as:

```
mychisq = 0
for (i in 1:1000) {
  newrefraction = sample(refraction)
  mychisq[i] = chisq.test(table(newrefraction, light))$statistic
}
```

- Use the [Analyzing Two-way Tables](#) applet. Copy and paste the two-way table (without totals and using only one word variable names) into the Sample Data box and press **Use Table**. Make sure the lighting condition is being used as the explanatory variable or press the **(explanatory, response)** button. Check the **Show Shuffle Options** box, enter a large number of shuffles, and press **Shuffle**. Use the pull-down menu to set the **Statistic** choice to the Chi-square ( $\chi^2$ ) statistic. Check the box to **Overlay Chi-square** distribution. You can also check the **Show  $\chi^2$  output** box to see the cell contributions.

(j) Does the randomization distribution appear to be well modeled by a chi-square distribution with  $df = 4$ ?

(k) Compute an empirical p-value based on the simulated chi-square values and compare it to the p-value calculated earlier. How do the p-values compare?

(l) If using R or Minitab, compute the p-value for each simulated chi-square value.

- **In R:** Use iscamchisqprob in your script

Describe the shape, center, and variability of the distribution of these p-values. Explain why this distribution makes sense. [Hint: Remember what assumption is behind the simulation.]

## Summary of Chi-Square Test of Association

Numerical and graphical summaries: conditional proportions and segmented bar graphs

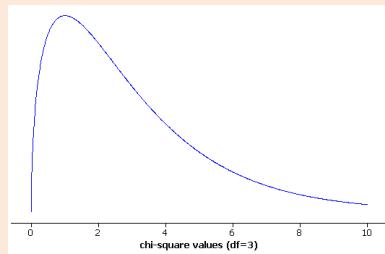
If the data are a random sample from a larger population and the average expected cell count is at least 5 (and all are at least 1) and you have classified each observational unit according to two categorical variables, then the hypotheses:

- H<sub>0</sub>: no association between *variable 1* and *variable 2*  
 vs. H<sub>a</sub>: there is an association between *variable 1* and *variable 2*

can be tested using the chi-square test statistic

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}.$$

The (upper-tail) p-value is calculated from the chi-square distribution with  $(r - 1)(c - 1)$  degrees of freedom.



### In R

- Create a matrix of the observed counts and use `chisq.test(matrix)`  
 To access the expected counts and (unsquared) chi-square contributions use `chisq.test(matrix)$expected` and `chisq.test(matrix)$residuals`.

### In Analyzing Two-way Tables applet

- You can paste in raw data or the two way table. Remember to use one word variable names and categories.
- Check the **Show Table** and **Show  $\chi^2$  output** boxes.

### Investigation 5.3: Newspaper Credibility Decline

With the proliferation of the Internet and 24-hour cable news outlets, it has become much easier for people to hear much more information, much more quickly. However, this has led to speculation that news organizations attempt to convey information before it has been properly verified in an effort to feed our impatience. *USA Today* reported that newspapers appear to be losing credibility over time (March, 2004). It cited a nationwide sample of 1,002 adults, 18 years or older, interviewed via telephone (under the direction of Princeton Survey Research Associates) during the period May 6-16, 2002. One of the questions asked was

Please rate how much you think you can BELIEVE each organization on a scale of 4 to 1. On this four point scale, “4” means you can believe all or most of what the organization says. “1” means you believe almost nothing of what they say. How would you rate the believability of (READ ITEM: ROTATE LIST) on this scale of 4 to 1?

The interviewer then asked this question for several different news organizations (e.g., *USA Today*, NPR, MSNBC).

- (a) Why is it important for the interviewer to rotate the list of organizations?

When asked about “The daily newspaper you are most familiar with,” the percentage distribution of the 1002 responses in May, 2002 was:

| <b>Believe all or almost all – 4</b> | <b>3</b> | <b>2</b> | <b>Believes almost nothing – 1</b> | <b>Can’t Rate</b> |
|--------------------------------------|----------|----------|------------------------------------|-------------------|
| 20%                                  | 39%      | 25%      | 9%                                 | 7%                |

A similar study conducted in May, 1998 yielded the following results (981 responses).

| <b>Believe – 4</b> | <b>3</b> | <b>2</b> | <b>Cannot Believe – 1</b> | <b>Can’t Rate</b> |
|--------------------|----------|----------|---------------------------|-------------------|
| 27%                | 36%      | 24%      | 7%                        | 6%                |

- (b) Create a two-way table of *counts* to analyze the association between time and how people rate the believability of the daily newspaper they are most familiar with, *among those who felt they could rate their daily newspaper*. In other words, eliminate the “can’t rate” category from consideration. Also remember to use the “explanatory variable” as the column variable and that the observed counts *do* need to be integers.

**Discussion:** If we have independent random samples from several populations as in Investigation 5.1, but the response variable has more than 2 categories ( $r > 2$ ), we can again extend the chi-square procedure. The expected counts and test statistic will be calculated exactly the same way as in Investigation 5.2. The chi-square distribution with  $(r - 1) \times (c - 1)$  degrees of freedom will be valid with the same technical condition on sample size and the null hypothesis is that the distributions of the response variable are the same in all  $I$  populations. The alternative is that the  $I$  distributions are not all the same.

You should notice that the mechanics are exactly the same whether you are testing homogeneity of proportions, comparing the distributions across several populations, or examining the association between two categorical variables. The difference lies in how the data were collected and therefore in the scope of conclusions that can be drawn. In fact, these procedures are sometimes all phrased in terms of the association between variables, because this is mathematically equivalent to having the same conditional distributions. Again, you need to consider the implications of whether or not the distribution of a variable was determined in advance by the researchers.

- (c) State the null and alternative hypotheses, check the technical conditions, and carry out a chi-square test to assess whether the sample data provide strong evidence that the distribution of responses differs between the two years. State your conclusion in context.

### Comparison to two-sample z-test

If we collapse the above table to focus on the “largely believable” (3 or 4) and “not largely believable” (1 or 2) breakdown, we obtain:

|                        | 2002 sample | 1998 sample | Total |
|------------------------|-------------|-------------|-------|
| Largely believable     | 591         | 618         | 1209  |
| Not largely believable | 341         | 304         | 645   |
| Total                  | 932         | 922         | 1854  |

- (d) Carry out a chi-square test to decide whether the difference in the population proportions giving a largely believable rating differed significantly between the two years.

- (e) Now carry out a *two-sided* two-proportion *z*-test for this table. Report the test statistic and p-value. How do the p-values compare? What do you think is the relationship between the test statistics?

**Discussion:** The chi-square procedure can be used to compare two or more population proportions. When there are only two populations, the procedure is equivalent to a two-sided  $z$ -test for proportions from Chapter 3. The chi-square statistic is equal to  $z^2$  and the chi-square p-value is equal to the two-sided p-value for the two-sample  $z$ -test. If the alternative hypothesis is two-sided, you can use either procedure. If the alternative hypothesis is one-sided, then you should carry out the two-sample  $z$ -test to obtain the one-sided p-value. If there are more than two populations, then you must use the chi-square procedure, which will only assess whether or not at least one population proportion differs from the others.

For a  $2 \times 2$  two-way table, another alternative is Fisher's Exact Test as you learned in Chapter 3. It is always appropriate to carry out Fisher's Exact Test, though it may be a bit less convenient with huge sample sizes (which becomes less of an issue with each new computer chip).

However, of all three of these procedures, the two-sample  $z$ -procedure is the only one that also enables us to calculate a confidence interval to estimate the magnitude of the difference in the two population proportions.

### Practice Problem 5.3

In February of 1993, NBC News admitted that it staged the explosion of a General Motors truck during a segment of the program "Dateline NBC" in November of 1992. The segment included crash footage that explosively showed how the gas tanks of certain old GM trucks could catch fire in a sideways collision. In a nationwide poll of adults (*Times Mirror News Interest Index*) conducted in August, 1989, 1507 respondents gave NBC news the following believability ratings.

| Believe – 4 | 3   | 2   | Cannot believe – 1 | Can't rate |
|-------------|-----|-----|--------------------|------------|
| 32%         | 47% | 14% | 2%                 | 5%         |

The same poll conducted Feb. 12-27, 1993 saw 2001 respondents give the following results:

| Believe – 4 | 3   | 2   | Cannot believe – 1 | Can't rate |
|-------------|-----|-----|--------------------|------------|
| 31%         | 42% | 18% | 6%                 | 3%         |

- (a) Is the difference in believability ratings statistically significant between these two years? (Is this a test of homogeneity of proportions or a test of independence? Explain.)
- (b) Is this convincing evidence that the General Motors explosion caused a decrease in the believability of NBC News?

## Section 5.1 Summary

This section introduced you to the *chi-square test*. You saw that this test arises in three different settings. One of these settings is comparing the proportions of success across two or more populations, when independent random samples are taken from the populations. Another setting is comparing the entire distributions of a categorical response variable (which could include more than just two categories) across two or more populations, again assuming independent random samples from each population. The chi-square test in these two scenarios is testing *homogeneity of proportions*. The final setting to which the chi-square test applies is testing *independence* between two categorical variables when a random sample is drawn from a population and the observational units are classified according to two categorical variables.

You have found that the same chi-square test procedure applies to all of these settings. The test statistic is based on comparing observed counts in the two-way table to what counts would be *expected* if the null hypothesis were true. You used simulation to investigate the sampling distribution of the chi-square test statistic, and studied the conditions under which that sampling/randomization distribution can be well-approximated by the chi-squared probability distribution, providing an approximation of the test's *p*-value. Finally, you learned that when the test indicates a statistically significant result, you can further analyze the contributions of the individual cells in the table to the calculation of that test statistic to learn more about the nature of the association. Example 5.1 presents an application of the chi-square procedure.

## SECTION 2: COMPARING SEVERAL POPULATION MEANS

In the previous section, you learned the chi-square test for comparing proportions among two or more groups. One key point was that there are advantages to an overall procedure that compares the proportions simultaneously and controls the overall Type I error rate. When there are only two groups, this procedure was equivalent to the two-sided two-sample z-test for proportions. You will see a similar approach in this section, addressing the issue of comparing two or more population *means*. The technique you learn here will be used both for comparing population means based on independent random samples and also for assessing whether there is a treatment effect based on data from a randomized experiment.

### Investigation 5.4: Disability Discrimination

The U.S. Vocational Rehabilitation Act of 1973 prohibited discrimination against people with physical disabilities. The act defined a disabled person as any individual who had a physical or mental impairment that limits the person's major life activities. In 1984, disabled individuals in the labor force had an unemployment rate of 7% compared to 4.5% in the non-impaired labor force.

Researchers conducted a study in the 1980s that examined whether physical disabilities affect people's perceptions of employment qualifications (Cesare, Tannenbaum, & Dalessio, 1990). The researchers prepared videotaped job interviews, using the same actors and script each time. The only difference was that the job applicant appeared with different disabilities:

- No disability
- Leg amputation
- Crutches
- Hearing impairment
- Wheelchair confinement

Seventy undergraduate students were randomly assigned to view one of the videotapes, and they were then asked to rate the candidate's qualifications on a 1-10 scale. The research question is whether subjects tend to evaluate qualifications differently depending on the applicant's disability.

(a) Identify the observational units, explanatory variable, and response variable in this study. Identify each variable as quantitative or categorical. Is this an observational study or an experiment? Explain.

Observational units:

Explanatory:

Type:

Response:

Type:

Type of study:

(b) The mean applicant qualification scores for the five groups are amputee 4.429, crutches 5.921, hearing 4.050, none 4.900, and wheelchair 5.343. What additional pieces of information do you need in order to decide whether these sample means are significantly different (more than you might expect by the random assignment process alone)?

(c) State a null and an alternative hypothesis to reflect the researchers' conjecture.

Null hypothesis:

Alternative hypothesis:

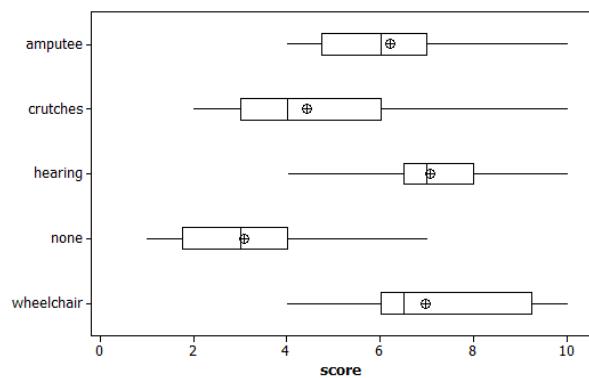
(d) Explain what a Type I error and a Type II error represent in this context.

Type I error:

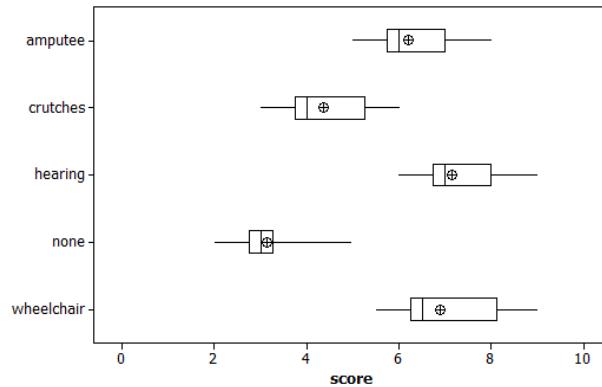
Type II error:

(e) Consider the following sets of boxplots. What is the same and what is different about the distributions displayed by these boxplots? In which group do you believe the evidence will be stronger that at least one population mean differs from the others? Explain.

A:



B:



Same:

Different:

Stronger evidence (A or B): Explain:

**Discussion:** In the above boxplots, the sample mean scores are roughly the same for graphs A and B (6.2, 4.4, 7.1, 3.1, 7.0)), but the variability within each group is much larger for the distributions in graph A. This larger amount of “within-sample” variation makes the differences in sample means seem not as extreme compared to graph B. Graph B displays less within-sample variability and so the differences in the sample means appear more extreme and provide more evidence that the crutches scores and the hearing scores did not come from the same population with the same population mean. So, our goal in comparing the sample means will be to decide whether the differences in the sample means are larger than what we would expect by random chance, where the amount of within-sample variation will give us an indication as how much we expect the response values to vary “by chance.”

- (f) The actual data are stored in [DisabilityEmployment.txt](#). Produce parallel boxplots and descriptive statistics of the qualification scores among the five groups. Record the descriptive statistics in the table below.

|                         | None | Amputee | Crutches | Hearing | Wheelchair |
|-------------------------|------|---------|----------|---------|------------|
| <b>Sample size</b>      |      |         |          |         |            |
| <b>Sample mean</b>      |      |         |          |         |            |
| <b>Sample std. dev.</b> |      |         |          |         |            |

Do these data suggest evidence of a difference among the population means? Write a paragraph supporting your statements.

Just as the chi-square test statistic provides a single number to compare how different sample proportions are across several groups, we need a test statistic that simultaneously measures the differences in sample *means* among several groups. As noted above, we will do this by considering the variation in the sample means from the overall mean, and assessing whether that variation is much larger than would be expected based on the chance variation exhibited within the samples. (See the Applet Exploration to explore another possible statistic.)

- (g) What is the overall mean applicant qualification rating assigned by the 70 students?

- (h) If we were to treat the five sample means as five observations, calculate the standard deviation of these five values. Then square the standard deviation to calculate the *variance*.

- (i) Is it reasonable to allow each of these sample means to have the same relative contribution to our overall measure of variability between group means? Explain.

Our measure of the *variability between groups* or “treatment effect” will be the *weighted variance* across the groups where the weights are the sample sizes.

$$\frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2 + n_5(\bar{x}_5 - \bar{x})^2}{5-1}$$

- (j) In this case, the sample sizes were equal to 14 in each group, so you can simply multiply your answer to (h) by 14 to measure the variability between groups:

Between-group variability =

In order to assess whether the sample group means vary enough to be considered statistically significant, we also need to consider the variability of qualification scores *within* each group. This will allow us to decide whether our deviations are larger than we would expect by chance. To take into account information from all five groups, we will calculate the *pooled variance* across the groups (similar to what was done in Investigation 3.5).

- (k) Calculate the average variance across the five groups. [Hint: Square each group’s standard deviation to obtain its variance, and then take the average of those five values.]

When the sample sizes are not equal, the more general formula for this pooled variance, weighted by sample size, is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2 + (n_5 - 1)s_5^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) + (n_5 - 1)}.$$

This provides an overall estimate of the “within-group” variability, which simplifies to the average of the group variances when the sample sizes are equal.

- (l) In calculating a pooled variance, you are implicitly assuming that the variability is the same across the groups. Explain how the sample statistics reveal that this is a reasonable belief for these data.

Our test statistic will consider the *ratio* of the variability between groups to the variability within groups.

- (m) Calculate the ratio of your result from (j) divided by your result from (k).

- (n) What is the smallest possible value that this ratio could ever assume? The largest?

(o) What types of values (large, small, positive, negative) will this ratio have when the null hypothesis is false, that is, when the population means are not all equal?

(p) Explain how we could simulate a randomization test to approximate the distribution of this test statistic.

(q) Use technology to simulate such a randomization test by randomly assigning the 70 rating scores to the five different disability groups (so that we are assuming no genuine treatment effect).

- **In R:** Modify the code used in Investigation 4.2. Remember to initialize the vector for storing the new statistic. Use `tapply` to keep track of group statistics. For example:  
`> means=tapply(randomized_score, disability, mean)`  
`> sds=tapply(randomized_score, disability, sd)`
- **In the Comparing Groups (Quantitative) applet:** Copy and paste the raw data, being to clarify which is the explanatory variable, and press **Use Data**. Check the **Show Shuffle Options** box, enter a larger number of shuffles, and press **Shuffle Responses**. Use the **Statistic** pull-down menu to select the **F-statistic**.

(r) Calculate an empirical p-value for the ratio you calculated in (m). What conclusion would you draw based on this p-value?

*Probability Result:* For large samples, this sampling distribution is well modeled by an *F distribution* with parameters *number of groups – 1* and *overall sample size – number of groups*, the degrees of freedom of the numerator and denominator respectively.

### Terminology Detour

We will compare  $I$  group means, where each group has  $n_i$  observations. The overall sample size will be denoted by  $N = \sum n_i$ .

$H_0$ : There is no treatment effect or  $H_0: \mu_1 = \dots = \mu_I$

$H_a$ : There is a treatment effect or  $H_a$ : at least one  $\mu_i$  differs from the rest

The between-group variability will be measured by looking at the sum of the squared deviations of the group means to the overall mean,  $\bar{x}$ . Each group mean is weighted by the sample size of that group. We will refer to this quantity as the “sum of squares for treatment,” SST.

$$SST = \sum_{i=1}^I n_i (\bar{x}_i - \bar{x})^2$$

We will then “average” these values by considering how many groups were involved. This quantity will be referred to as the “mean square for treatments.”

$$MST = \frac{SST}{(I-1)}$$

Note, if we fix the overall mean, once we know  $I-1$  of the group means, the value of the  $i^{\text{th}}$  mean is determined. So the degrees of freedom of this quantity is  $I-1$ .

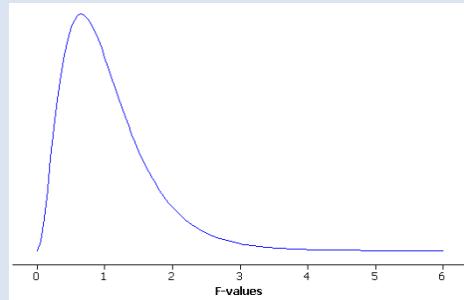
The within-group variability will be measured by the pooled variance. In general, each term will be weighted by the sample size of that group. We will again divide by an indication of the overall sample size across the groups. We will refer to this quantity as the “mean squares for error,” MSE.

$$MSE = \frac{\sum_{i=1}^I (n_i - 1) s_i^2}{N-I} \text{ which has } N-I \text{ degrees of freedom.}$$

The test statistic is then the ratio of these “mean square” quantities:

$$F = \frac{MST}{MSE}$$

When the null hypothesis is true, this test statistic should be close to 1. So larger values of  $F$  provide evidence against the null hypothesis. The corresponding p-value comes from a probability distribution called the *F distribution* with  $I-1$  and  $N-I$  degrees of freedom.



We will use this *F* distribution to approximate both the sampling distribution of this test statistic in repeated samples from the same population ( $H_0: \mu_1 = \dots = \mu_I$ ) and the randomization distribution for a randomized experiment ( $H_0$ : no treatment effect) as long as the technical conditions (see below) are met.

Because we are focusing on the variance of group means, this procedure is termed Analysis of Variance (ANOVA). With one explanatory variable, this is called *one-way ANOVA*.

## Technology Detour – ANOVA

### In R

- If you have the data in a response vector and an explanatory vector, you can use `summary(aov(response~explanatory))`

This will output the Mean Square values in the Mean Sq column, first for the treatment group (the *disability* row) and then for the error term (Residuals).

### In [Comparing Groups \(Quantitative\)](#) applet

- Paste data into Sample data box, being sure to clarify which is the explanatory variable. Press **Use Data**.
- Check the **Show ANOVA Table** box.

(s) Verify that the *F*-statistic is then equal to the ratio of these two Mean Square values. Also verify the df values (degrees of freedom) for each row. Report the p-value.

MST:

MSE:

*F*-statistic:

df numerator:

df denominator:

p-value:

## Technical Conditions

There are several technical conditions required for this randomization distribution to be well modeled by the *F* distribution:

- The distribution for each group comes from a normal population.
- The population standard deviation is the same for all the groups.
- The observations are independent.

When using this ANOVA *F*-test to compare several population means, we will check each condition as follows:

- The normal probability plot (or dotplot or histogram) for each sample's responses is reasonably well-behaved.
- The ratio of the largest standard deviation to the smallest standard deviation is at most 2.
- The samples are independent random samples from each population.

For simplicity, we will apply the same checks for a randomized experiment as well (with the last condition being met if the treatments are randomly assigned). If the first two conditions are not met, then suitable transformations may be useful.

(t) Is there evidence of non-normality in these sample data? Calculate the ratio of the largest to smallest standard deviation; is this less than 2? (In the applet, check the box to Overlay *F* distribution to help assess the fit.)

- (u) Write a paragraph summarizing your conclusions for this study. Be sure to address issues of causation and generalizability as well as statistical significance.

### Study Conclusions

The distributions of qualification scores in the five treatment groups look reasonably symmetric with similar standard deviations, so it is appropriate to apply the Analysis of Variance procedure. There is moderate evidence that the mean qualification ratings differ depending on the type of disability ( $p$ -value = 0.030). Descriptively, the candidates with crutches appear to have higher ratings on average, and the candidates with hearing impairments tend to have slightly lower ratings. (Other follow-up procedures could be used to determine which group means differ significantly from which others.) This was a randomized experiment, so we can attribute these differences in qualification scores to the disability shown. But we must be cautious about considering the students in this study to be representative of a larger population, particularly a population of employers who make actual hiring decisions.

### Practice Problem 5.4A

Reconsider the boxplots from part (e). Explain how they compare and how this will affect the calculation of the  $F$ -statistic and lead to different  $p$ -values. Explain based on how the  $F$ -statistic is calculated which set of boxplots will have the smaller  $p$ -value.

### Practice Problem 5.4B

Lifetimes of notable people in nine different occupational categories were gathered from *The World Almanac and Book of Facts*.

- (a) How many different *pairs* of occupations can be considered?
- (b) Explain why conducting an ANOVA is different from conducting separate two-sample  $t$ -tests on all possible pairs of occupations. Include an explanation for why it would be inappropriate to do all of those separate two-sample  $t$ -tests.
- (c) Would rejecting the null hypothesis in ANOVA allow you to conclude that every occupation being studied has a different population mean lifetime from every other occupation? Explain.

## Applet Exploration: Randomization Test for ANOVA

An alternative to using ANOVA for comparing a quantitative response variable across multiple groups is to conduct a randomization test. As you have encountered with other scenarios, the randomization test involves

- shuffling the observed values for one of the variables, say the response variable,
- recalculating the value of the statistic for the shuffled data,
- repeating that a large number of times to see the simulated null distribution of the statistic, and then
- approximating a p-value based on where the observed value of the statistic falls in the simulated null distribution.

One advantage of the randomization test is that it can be performed on any reasonable statistic that measures how much the response variable differs across the groups. You are no longer limited to using the ANOVA  $F$ -statistic, which might not have been the first statistic that would have occurred to you.

Reconsider the experiment from Investigation 5.4 about disability discrimination.

(a) State again the null hypothesis to be tested.

(b) Report again the sample mean qualification score for each group.

| Group       | None | Amputee | Crutches | Hearing | Wheelchair |
|-------------|------|---------|----------|---------|------------|
| Sample mean |      |         |          |         |            |

(c) Suggest a statistic for measuring how far apart these group means are. [Hints: Your statistic should be a formula that produces a single number. Propose a statistic that is simpler and easier to calculate than the  $F$ -statistic.]

One simple but reasonable statistic is to calculate the average of all the absolute differences in sample means between pairs of groups. This statistic is called MAD (mean of absolute differences).

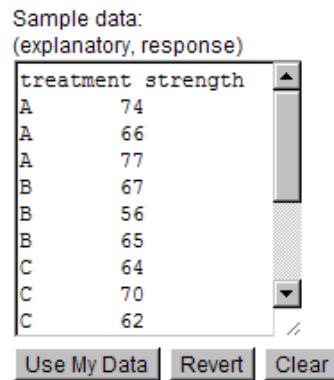
(d) How many *pairs* of groups are there in this study? [Hint: Think combinations.]

(e) Why is it important to take the absolute values of the pairwise differences in group averages before summing them? [Hint: Trying summing this without the absolute values.]

(f) Calculate the value of the MAD statistic for the disability discrimination.

After we calculate the value of the MAD statistic for the sample data, the next question is whether this value is larger than would be expected by random chance if there were no treatment effect (i.e., if the response values would have turned out the same regardless of which explanatory group the person had been assigned to). As we have done before in other situations, we can again investigate this question by simulating the distribution of the MAD statistic under the null hypothesis.

- (g) Open the [Comparing Groups \(Quantitative\)](#) applet to simulate a null distribution of these MAD statistics. Open the [DisabilityEmployment.txt](#) data file and paste the two columns (with column names) into the applet; then press **Use Data**. Verify that the sample means agree with your answer to (a). Then report the value that the applet provides for the MAD statistic, and confirm that this matches your answer to (f).



- (h) Check the **Show Shuffle Options** box in the applet. Make sure that the radio button for the MAD statistic is checked. Press **Shuffle Responses**. Notice that the new dotplots and statistics are produced for the shuffled data. Report the value of the MAD statistic for the shuffled data. Is this value more extreme than the observed value of the MAD statistic from the experimental data? (How are you deciding?)

- (i) Press **Shuffle Responses** four more times. Record the four new values of the MAD statistic for the shuffled data. Are these all the same? Are any of them more extreme than the observed value of the MAD statistic from the experimental data?

- (j) Now enter 995 for the **Number of Shuffles** and press **Shuffle Responses**, producing a total of 1000 shuffles. Describe the shape of the null distribution of MAD statistic values. Is it symmetric or skewed? If it is skewed, in which direction is it skewed?

- (k) Is the null distribution centered at zero? Explain why your answer makes sense.

- (l) In order to calculate an empirical p-value, you will count how many of the simulated MAD statistics are equal to \_\_\_\_\_ or <larger or smaller>.

- (m) In the **Count Samples** box enter the observed value of the MAD statistic and press **Count**. Report the empirical p-value.

- (n) To obtain a more accurate approximation to the p-value, generate 9000 more random shuffles. Report the new empirical p-value.
- (o) Interpret this p-value: It is the probability of obtaining a MAD statistic of \_\_\_\_\_ or \_\_\_\_\_, assuming that \_\_\_\_\_.
- (p) Based on the p-value, summarize the strength of evidence that the sample data provide against the null hypothesis, in the context of this study. Also explain the reasoning process behind your conclusion.

The MAD statistic is fairly simple and intuitive, but it is not widely used. The *F*-statistic that you studied earlier is far more commonly used, and the *F*-statistic has the advantage of taking into account not only the group means but also within-group variability. As a result, the *F*-statistic is more powerful than the MAD statistic.

- (q) Explain what it means to say that the *F*-statistic is more powerful than the MAD statistic.

The applet conducts a randomization test based on the *F*-statistic as well as the MAD statistic.

- (r) Use the pull-down **Statistic** menu to the **F-statistic**, and the null distribution changes from the MAD statistic to the *F*-statistic. Report the observed value of the *F*-statistic, which appears on the left side of the applet.
- (s) Use the simulation results, and the observed value of the *F*-statistic, to determine an empirical p-value.
- (t) Check the box by **Show ANOVA Table**, and the applet reports more output including the p-value based on the *F*-distribution. Report this p-value. Is the empirical p-value from (s) close to this theoretical p-value? Is the theoretical *F*-distribution, which is now overlaid on the simulated null distribution, a reasonable match?
- (u) You have now calculated a p-value in three different ways, two using simulations of randomization tests (with the MAD and *F*-statistics) methods and one based on the theoretical *F*-distribution. Are the p-values all similar? Do they all lead to the same conclusion about the strength of evidence that handicap type affects perception of qualification?

### Investigation 5.5: Restaurant Spending and Music

A British study (North, Shilcock, & Hargreaves, 2003) examined whether the type of background music playing in a restaurant affected the amount of money that diners spent on their meals. The researchers asked a restaurant to alternate classical music, popular music, and silence on successive nights over 18 days. The following summary statistics were reported:

|                    | <b>Classical music</b> | <b>Pop music</b> | <b>No Music</b> |           |
|--------------------|------------------------|------------------|-----------------|-----------|
| <b>Mean</b>        | £24.13                 | £21.91           | £21.70          |           |
| <b>SD</b>          | £2.243                 | £2.627           | £3.332          |           |
| <b>Sample size</b> | $n_1 = 120$            | $n_2 = 142$      | $n_3 = 131$     | $N = 393$ |

- (a) Based on these summary statistics, calculate the overall (weighted) mean amount spent and the pooled standard deviation. Check the consistency of your calculations with the summary statistics reported above.

Overall (weighted) mean:

Pooled standard deviation:

- (b) State the null and alternative hypotheses corresponding to the researchers' conjecture.

- (c) Based on the above summary statistics and part (a), calculate the  $F$ -statistic (by hand) and then use technology to determine the p-value (sketch and shade the corresponding randomization distribution):

- In R: `pf(x, df1, df2, lower.tail=FALSE)`

- (d) What additional information is necessary before you can continue to draw a conclusion from this p-value?

## Study Conclusions

In this study, a larger average amount was spent while classical music was playing. With such a large test statistic and such a small p-value ( $F = 31.48$ , p-value  $\approx 0$ ), we have very strong evidence that the observed differences in the sample means could not have arisen by chance alone. However, we need to assess the technical conditions before we could know whether this inferential procedure is valid with these data. Although the equal standard deviation assumption seems reasonable (because  $3.332/2.243 < 2$ ), we do not have the sample data to examine the shapes of the sample distributions. Still, the sample sizes are large and the  $F$  procedure is fairly robust to departures from the normality condition. However, we should *not* consider these observations to be *independent* within each group, because the treatments were assigned to the evenings, not to the individual diners. In this sense, we only have a few replications per evening and ANOVA is not the appropriate analysis. Descriptively it appears that the average amount spent is larger when classical music is playing, but because of the confounding with evening, it's possible that classical music was played on certain days of the week and that that lead to more spending. This was a randomized experiment, so if it weren't for the concern about independence, we would have been able to draw a cause-and-effect conclusion from the type of music played and the amount spent. We would want to be very cautious in generalizing these results to other restaurants and even other times of the year.

## Applet Exploration: Exploring ANOVA

You will now explore the effects of such factors as the size of the difference in the population means, the overall population standard deviations, and the sample sizes on the  $F$ -test statistic and p-value.

Open the [Simulating ANOVA Tables](#) applet.  
The sliders and text boxes should specify:

- 23 as each population mean
- 120, 142, and 131 as the sample sizes
- 3 as the population standard deviation.

### Simulating ANOVA Tables

|              |                                             |
|--------------|---------------------------------------------|
| $\mu_1 = 23$ | $n_1 = 120$                                 |
| $\mu_2 = 23$ | $n_2 = 142$                                 |
| $\mu_3 = 23$ | $n_3 = 131$                                 |
| $\sigma = 3$ | <input type="button" value="Draw Samples"/> |

(e) Press the **Draw Samples** button. A sample will be selected from each population. What  $F$ -statistic and p-value did you obtain?

(f) Press the **Draw Samples** button again. Did you obtain the same  $F$ -statistic and p-value? Why not?

(g) Press the **Draw Samples** button 10-20 more times, watching how the boxplots,  $F$ , and p-values change. Did you ever obtain a p-value below 0.05? Is this possible? Would it be surprising? Explain.

(h) Change the value of  $\mu_1$  to **24**. Press the **Draw Samples** button 10-20 times. Do the resulting p-values tend to be larger or smaller than in (g)? Explain why this makes sense.

(i) Change each of the **sample sizes** to **20** and press **Draw Samples** 10-20 times. Do the resulting p-values tend to be larger or smaller than in (h)? Explain why this makes sense.

(j) Press **Draw Samples** until you have a p-value  $< 0.3$ . Now change the value of  $\sigma$  to 7. (Drag the slider to decrease the value in increments of 0.1, or edit the orange value, watching how the p-value changes.) How does this affect the magnitude of the p-value? Explain why this relationship makes sense.

(k) How does the p-value generally change if you now continue to increase the value of  $\mu_1$ ? Explain why this relationship makes sense.

**Discussion:** If the null hypothesis is true, then the p-value should vary uniformly between 0 and 1. In this case, for example, the p-value will be less than 0.05 in 5% of all random samples, so 5% of samples would lead you to reject the null hypothesis even when it is true. However, when there truly is a difference in the population means, our ability to detect that difference based on sample data is affected by several factors:

- When the population means are further apart, the p-value is smaller.
- When the within-group variability is larger, the p-value is larger.
- When the sample sizes are larger, and there is a difference among the population means, then the p-value is smaller.

The p-value of any particular study is in essence random, so we need to remember the Type I and Type II errors that we could be making. Committing a Type I error with Analysis of Variance indicates that we concluded that the population means differ when they really don't differ. Type II error indicates that we failed to conclude that at least one population mean differs when the population means are in actuality not all equal.

**Practice Problem 5.5A**

Another way to analyze the data for the trial of Dr. Spock is to look at the percentages of women on the different venires and determine whether the mean percentage of women is equal across the seven judges. Below are the percentages of women on the venires for a recent sample from each of the judges. These data are below and in [SpockPers.txt](#).

| Judge 1 | Judge 2 | Judge 3 | Judge 4 | Judge 5 | Judge 6 | Judge 7 |
|---------|---------|---------|---------|---------|---------|---------|
| 16.8    | 27.0    | 21.0    | 24.3    | 17.7    | 16.5    | 6.4     |
| 30.8    | 28.9    | 23.4    | 29.7    | 19.7    | 20.7    | 8.7     |
| 33.6    | 32.0    | 27.5    | 21.5    | 23.5    |         | 13.3    |
| 40.5    | 32.7    | 27.5    | 27.9    | 26.4    |         | 13.6    |
| 48.9    | 35.5    | 30.5    | 34.8    | 26.7    |         | 15.0    |
| 45.6    | 31.9    | 40.2    | 29.5    |         |         | 17.7    |
| 32.5    | 29.8    |         |         |         |         | 18.6    |
| 33.8    | 31.9    |         |         |         |         | 23.1    |
| 33.8    | 36.2    |         |         |         |         | 15.2    |

- (a) Explain what information we learn from analyzing the data this way that we did not see when we carried out the Chi-Square test on the overall proportion of women for each judge. Why might this information be useful?
- (b) Produce numerical and graphical summaries to compare the percentages across the seven judges.
- (c) Carry out an ANOVA to test whether at least one judge has a different mean percentage. Did you state the null and alternative hypotheses in terms of population parameters or in terms of treatment effects?
- (d) Comment on whether you believe the technical conditions for this procedure are met.

**Practice Problem 5.5B**

Recall the Disability Discrimination study (Investigation 5.4). Suppose the researchers had been most interested in comparing the results for those in a wheelchair to those with leg amputation.

- (a) Carry out a two-sided two-sample [pooled t-test](#) (see Investigation 4.2: use the technology option that assumes the variances are equal) to assess whether there is a statistically significant difference in the average ratings assigned to these two groups.
- (b) Carry out an analysis of variance to assess whether there is a statistically significant difference in the average ratings assigned to these two groups.
- (c) How are the p-values from (a) and (b) related? How do you think the *t*-test statistic and the *F*-test statistic are related? Explain.
- (d) Suggest (in general) a situation where the two-sample *t*-procedure would be preferred and a situation where the ANOVA procedure would be preferred. [Hint: What would be true about the research question?]

## Section 5.2 Summary

Whereas the previous section dealt with comparing several groups on a categorical response variable, this one introduced you to comparing several groups on a quantitative response variable. You learned how to apply *analysis of variance* (ANOVA) to assess the significance of the differences among several sample/treatment means. The test procedure is based upon a probability distribution known as the *F*-distribution. You have seen that the same ANOVA procedure applies whether the data were gathered as independent random samples from several populations or from randomization of subjects to several treatment groups. (But like always, the scope of conclusions differs depending on how the data were collected.) You have also found that ANOVA results are affected by the variability in sample means, the variability within samples/groups, and the sample sizes. Like always, you have examined technical conditions that must be satisfied in order for this procedure to be valid; in addition to random sampling or random assignment, these conditions are that each group has a normal distribution and that the standard deviations are similar across groups. Example 5.2 presents an application of the ANOVA procedure.

### SECTION 3: RELATIONSHIPS BETWEEN QUANTITATIVE VARIABLES

In this section you will analyze data sets with two quantitative variables. The goal will be to describe the relationship between the variables. As always, you will start by learning some useful numerical and graphical techniques for summarizing the data. Then you will explore how to use a mathematical model of the relationship to make predictions of one variable from the other. In the next section you will then move on to inferential techniques based on simulated sampling and randomization distributions as well as a mathematical model.

#### Investigation 5.6: Cat Jumping

Evolutionary biologists are often interested in “form-function relationships” to help explain evolution history of say an animal species. Harris and Steudel (2002) investigated factors that are related to the jumping ability of domestic cats. Because jump ability and height are largely dependent on takeoff velocity, several traits were recorded for 18 healthy adult cats such as relative limb length, relative extensor muscle mass, body mass, fat mass relative to lean body mass, and the percentage of fast-twitch muscle fibers to see which might best explain *maximum takeoff velocity* (based on high-speed videos). In this investigation, you will examine the following data, also available in the file [CatJumping.txt](#):

| ID | Sex | Body mass (g) | Hind limb length (cm) | Muscle mass (g) | Percent body fat | Takeoff velocity (cm/sec) |
|----|-----|---------------|-----------------------|-----------------|------------------|---------------------------|
| A  | F   | 3640          | 29.1                  | 51.15           | 29               | 334.5                     |
| B  | F   | 2670          | 28.55                 | 46.05           | 17               | 387.3                     |
| C  | M   | 5600          | 31.74                 | 95.9            | 31               | 410.8                     |
| D  | F   | 4130          | 26.9                  | 55.65           | 39               | 318.6                     |
| E  | F   | 3020          | 26.11                 | 57.2            | 15               | 368.7                     |
| F  | F   | 2660          | 26.69                 | 48.67           | 11               | 358.8                     |
| G  | F   | 3240          | 26.74                 | 64.55           | 21               | 344.6                     |
| H  | M   | 5140          | 27.71                 | 78.8            | 35               | 324.6                     |
| I  | F   | 3690          | 25.47                 | 54.6            | 33               | 301.4                     |
| J  | F   | 3620          | 28.18                 | 55.5            | 15               | 331.8                     |
| K  | F   | 5310          | 28.45                 | 68.8            | 42               | 312.6                     |
| L  | M   | 5560          | 28.65                 | 79.8            | 37               | 316.8                     |
| M  | M   | 3970          | 29.82                 | 69.4            | 20               | 375.6                     |
| N  | F   | 3770          | 26.66                 | 60.25           | 26               | 372.4                     |
| O  | F   | 5100          | 27.84                 | 60.7            | 41               | 314.3                     |
| P  | F   | 2950          | 27.89                 | 55.65           | 25               | 367.5                     |
| Q  | M   | 7930          | 30.58                 | 98.95           | 48               | 286.3                     |
| R  | F   | 3550          | 28.06                 | 79.25           | 16               | 352.5                     |

- (a) Identify the observational units and the primary *response* variable of interest here. Also classify this variable as quantitative or categorical.

Observational units:

Response variable:

Type:

(b) Open the [CatJumping.txt](#) data file and produce numerical and graphical summaries of the *takeoff velocity* variable. Describe the distribution of takeoff velocities in this sample (shape, center, variability, unusual observations).

(c) Based on your analysis in (b), if you were going to randomly select a domestic cat, what is your best prediction of its takeoff velocity?

(d) Do you think there will be a relationship between a cat's takeoff velocity and its body mass? If so, do you think heavier cats will tend to have larger or smaller takeoff velocities than lighter cats?

We will need a new graphical summary to visually explore the relationship between two quantitative variables, the *scatterplot*.

### Technology Detour – Scatterplots

#### In R

```
> plot(bodymass, velocity)  or > plot(velocity~bodymass)  
> scatterplot(bodymass, velocity)      the scatterplot function is in the car package
```

(e) Describe the relationship between a cat's takeoff velocity and its body mass, as displayed in this scatterplot. Does this pattern confirm your expectation in (d)?

(f) Do any of these cats appear to be outliers in the sense that its pair of values (body mass, takeoff velocity) does not fit the pattern of the majority of cats? If so, identify the ID for that cat and describe what's different about this cat (in context).

### Terminology Detour

Scatterplots are useful for displaying the relationship between two quantitative variables. If one variable has been defined as the response variable and the other as the explanatory variable, we will put the response variable on the vertical axis and the explanatory variable along the horizontal axis.

In describing scatterplots you will describe the overall pattern between the two variables focusing primarily on three things:

- *Direction*: Is there a *positive association* (small values of  $y$  tend to occur with small values of  $x$  and large values of  $y$  tend to occur with large values of  $x$ ) or a *negative association* (small values of  $y$  tend to occur at large values of  $x$  and vice versa)?
- *Linearity*: Is the overall pattern in the scatterplot linear or not?
- *Strength*: How closely are the observations following the observed pattern?

The above scatterplot reveals a fairly strong, negative association between body mass and takeoff velocity, meaning that heavier cats tend to have a smaller takeoff velocity than larger cats. The relationship is somewhat linear but has a bit of a curved pattern. There is one outlier cat (cat C) with a very high takeoff velocity despite having a very large body mass.

(g) Now produce a scatterplot of takeoff velocity vs. percentage of body fat. Describe the association. Would you say that the association with velocity is stronger than with body mass? More or less linear?

(h) For the other two variables (hind limb length and muscle mass), would you expect to see a positive or negative association with takeoff velocity? Explain. Then look at scatterplots, and comment on whether the association is as you expected.

Now produce a coded scatterplot of takeoff velocity vs. body mass that uses different symbols for male and female cats:

### Technology Detour – Coded Scatterplots

#### In R

- Create the scatterplot as before, but pass a categorical variable as a color vector.

For example: > `plot(velocity~bodymass, col=sex)`

- (i) Based on this graph, do you notice any differences between male and female cats with regard to these variables? Explain.

### Study Conclusions

These researchers reported that variation in cat maximum takeoff velocity was significantly explained by both hind limb length (cats with longer limbs tended to have higher takeoff velocities) and fat mass relative to lean body mass (cats with lower fat mass tended to have higher takeoff velocities), but not to extensor muscle mass relative to lean mass or fast-twitch fiber content. They explained the “pervasive effect” of body mass by the increase in muscle work invested in increasing the center of mass potential energy as compared with kinetic energy during takeoff.

Later in this chapter you will learn how they determined the statistical significance of these relationships. First, we will examine a numerical measure of the strength of the association between two variables.

### Investigation 5.7: Drive for Show, Putt for Dough

Some have cited “Drive for show, putt for dough” as the oldest cliché in golf. The message is that the best way to improve one’s scoring average in golf is to focus on improving putting, as opposed to, say, distance off the initial drive, even though the latter usually garners more ooh’s and aah’s. To see whether this philosophy has merit, we need to examine whether there is a relationship between putting ability and overall scoring, and whether that relationship is stronger than the relationship between scoring average and driving distance. The file [golfers.txt](#) contains the 2004 statistics (through the Honda Classic on March 20) on the top 80 PGA golfers, downloaded from

<http://www.pgatour.com/stats/> on March 20, 2004. Three of the variables recorded include:

- *Scoring average*: A weighted average which takes the stroke average of the field into account. It is computed by adding a player’s total strokes to an adjustment, and dividing by the total rounds played. This average is subtracted from par to create an adjustment for each round. Keep in mind that in golf low scores, as measured by number of strokes, are better than high scores.
- *Driving distance*: Average number of yards per measured drive. These drives are measured on two holes per round, carefully selected to face in opposite directions to counteract the effects of wind. Drives are measured to the point where they come to rest, regardless of whether or not they hit the fairway.
- *Putting average*: On holes where the green is hit in regulation, the total number of putts is divided by the total holes played.

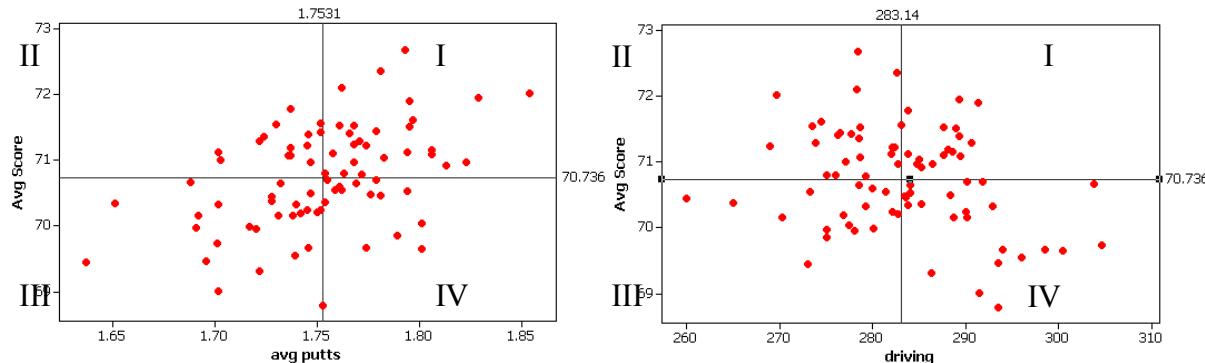
(a) Do you expect the relationship between scoring average and driving distance to be positive or negative? Explain.

(b) Do you expect the relationship between scoring average and putting average to be positive or negative? Explain.

(c) Open the data file and examine a scatterplot of *average score* vs. *driving distance* and a scatterplot of *average score* vs. *average putts*. Describe each scatterplot. Do the relationships confirm your expectations in (a) and (b)? Does one relationship appear to be stronger than the other? If so, which?

To further analyze these data, we need a numerical way of measuring the strength of the association between the two variables. We will do this with the *correlation coefficient*.

The following are the above scatterplots with the  $\bar{x}$  and  $\bar{y}$  lines superimposed.



- (d) For the *average score* vs. *average putts* scatterplot, in which quadrants are most of the points located? For the *average score* vs. *driving* scatterplot? Which scatterplot seems to have fewer points in the “non-aligned” quadrants?

**Definition:** A numerical measure of the strength of a linear relationship is the [correlation coefficient](#),  $r$ .

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

where  $\bar{x}$  and  $s_x$  are the mean and standard deviation of the explanatory variable, respectively, and  $\bar{y}$  and  $s_y$  are the mean and standard deviation of the response variable.

Notice that when a point  $(x_i, y_i)$  is in quadrant I or III, the term  $(x_i - \bar{x})(y_i - \bar{y})$  will be positive. When a point is in quadrant II or IV, this term is negative. Where there is a positive association, most of the points are in quadrants I and III, so the correlation coefficient is positive. Similarly, when there is a negative association, most of the points will be in quadrants II and IV, so the correlation coefficient turns out to be negative. The more observations that are in the “aligned” quadrants, and the closer the points fall to a straight line, the larger the value of the correlation coefficient.

- (e) Let  $x$  represent the average number of putts and  $y$  represent the average number of strokes. Use the definition of the correlation coefficient above to determine the measurement units of the correlation coefficient  $r$  in terms of putts and strokes.

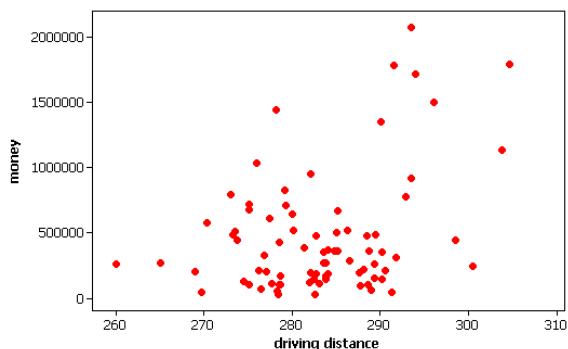
(f) If the correlation coefficient between two variables equals zero, what do you think the scatterplot will look like?

(g) Suppose we find the correlation coefficient of a variable with itself. Substitute  $x_i$  in for  $y_i$  (and so  $\bar{x}$  for  $\bar{y}$  and  $s_x$  for  $s_y$ ) in the above equation. Simplify. What is the correlation coefficient equal to?

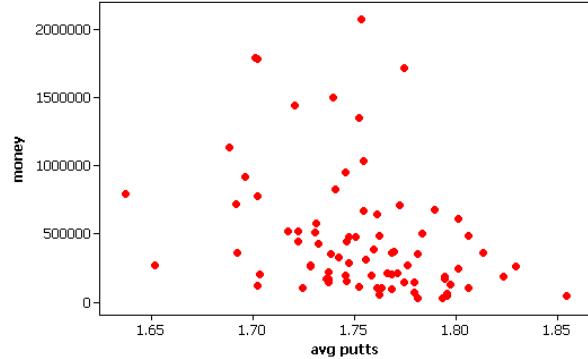
(h) Do you think the correlation coefficient will be a *resistant* measure of association? Explain.

(i) The following scatterplots display 7 pairs of variables for these golfers. Rank these graphs in order from strongest negative correlation to strongest positive correlation.

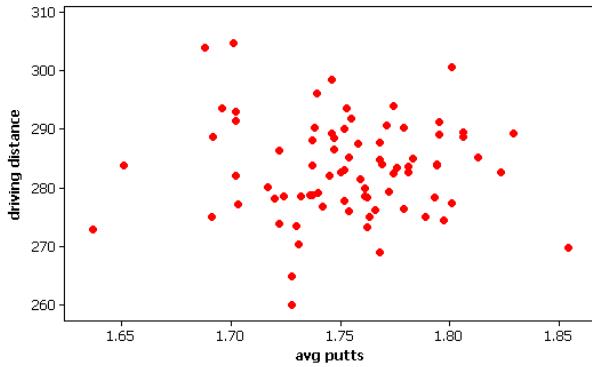
A:



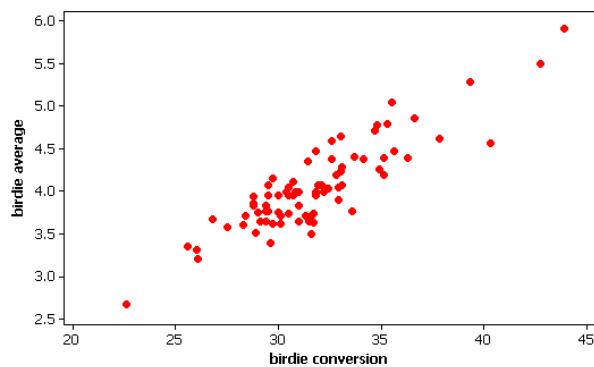
B:



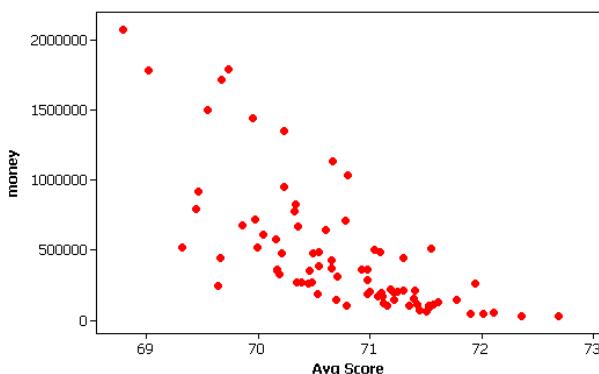
C:



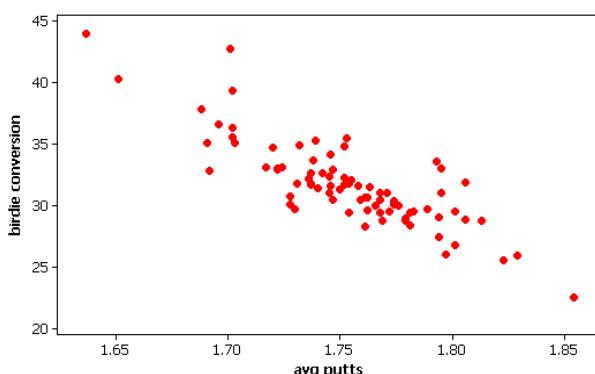
D:



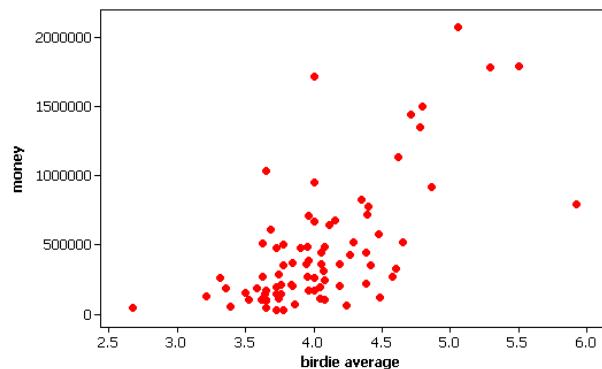
E:



F:



G:



Strongest negative:

Medium negative:

Weak negative:

No association:

Weak positive

Medium positive:

Strongest positive:

(j) Use technology to determine the correlation coefficient for each of the above scatterplots

- In R: > `cor(x, y)`

Record the values of these correlation coefficients below:

|                    |                                      |       |
|--------------------|--------------------------------------|-------|
| Strongest negative | birdie conversion and average puts   | _____ |
| Medium negative    | money and average score              | _____ |
| Weak negative      | money and average putts              | _____ |
| No association     | driving distance and average puts    | _____ |
| Weak positive      | money and driving distance           | _____ |
| Medium positive    | money and birdie average             | _____ |
| Strongest positive | birdie average and birdie conversion | _____ |

(k) Based on these correlation coefficient values and/or the definition/formula, what do you think is the largest value that  $r$  can assume? What is the smallest value? [Hint: It's not zero.]

Largest:

Smallest:

(l) If the association is negative, what values will  $r$  have? What if the association is positive?

(m) What does a correlation coefficient equal to zero signify?

(n) What does a correlation coefficient close to 1 or  $-1$  signify?

(o) Which has a stronger correlation coefficient with scoring average: driving distance or average putts? Does this support the cliché? Explain.

### Study Conclusions

The correlation coefficient for scoring average and average putts indicates a moderately strong positive linear association ( $r = 0.444$ ) whereas the correlation coefficient for scoring average and driving indicates a weaker negative association ( $r = -0.265$ ). This appears to support that putting performance is more strongly related to a PGA golfer's overall scoring average than the golfer's driving distance, as the cliché would suggest. We must keep in mind that these data are only for only the first 2.5 months of the season (when most golfers have played only around 6–8 events) and may not be representative of the scores and money earnings later in the year.

**Discussion:** The correlation coefficient,  $r$ , provides a measure of the strength and direction of the *linear* relationship between two variables. This is a unitless quantity which has the advantage that it will be invariant to changes in scale (if we started looking at money in British pounds instead of American dollars, our measure of the strength of the relationship will not change) or if we reverse which we call the explanatory and response variables. A value of  $r$  close to zero indicates that the variables do not have a strong linear relationship. However, this does not preclude them from having a very strong but non-linear relationship. A scatterplot should be examined before interpreting the value of  $r$ . If the relationship is not linear, there are alternative measures of the strength of the association that can be used or the variables can be *transformed* and the transformed variables analyzed instead. If the relationship is linear, a correlation coefficient close to 1 or  $-1$  indicates a very strong relationship. The values of  $r$  form a continuum: as the linear association becomes weaker, the value of  $r$  becomes closer to zero.

The correlation coefficient will always be a number between  $-1$  and  $1$ , inclusive. It will obtain the value of  $-1$  or  $1$  if the points fall along a perfect line (with negative or positive slope, respectively).

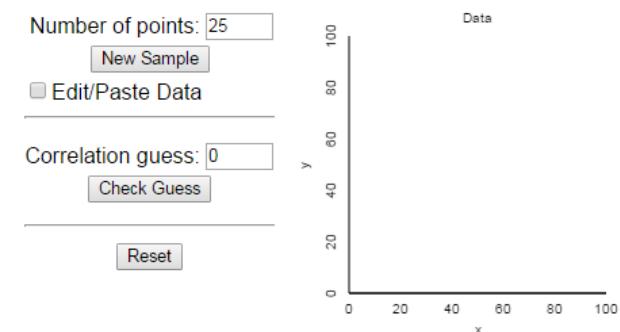
## Applet Exploration: Correlation Guessing Game

Open the [Guess the Correlation](#) applet.

- (a) Leave the number of points set to 25 and press **New Sample**. The applet will display a scatterplot. Guess the value of the correlation coefficient in this scatterplot and enter this guess into the “Correlation guess” box. Then press the **Check Guess** button.

How close was your guess? Were you surprised by the actual value? In what way? How might you guess differently the next time?

### Correlation Guessing Game



- (b) Press the **New Sample** button, enter your guess for the value of the correlation coefficient for this scatterplot, and press **Enter**. Describe how close you were and what adjustments you might make in your thinking.

- (c) Repeat this process for a total of 10 scatterplots (you do not need to record your results but you should try to learn from the reveal of the actual values as you make your guesses). Do you think your guessing ability improved by the last scatterplot? Explain.

- (d) Check the **Track Performance** box. The first graph is “Guess vs. Actual” graph with the ( $y = x$ ) line shown. Describe the behavior of the relationship between your guesses and the actual values of the correlation coefficients. Is there a strong correlation between your guesses and the actual values? Does this mean you are a good guesser? Explain.

- (e) The next graph is the “Error vs. Actual” graph. Describe the behavior of the relationship between your errors and the actual values of the correlation coefficients. Were some correlation coefficients easier for you to guess than others? Use this graph to justify your answer.

(f) The last graph is the “Error vs. Trial” graph. Describe the behavior of the relationship between your errors and the order in which you saw the graphs. Did your ability seem to improve over time? Use this graph to justify your answer.

(g) Suppose you guessed every value correctly; what would be the value of the correlation coefficient between your guesses and the actual correlations?

(h) Suppose each of your guesses was too high by 0.2 from the actual value of the correlation coefficient. What would be the value of the correlation coefficient between your guesses and the actual correlations?

(i) Does a correlation coefficient equal to 1 necessarily imply you are a good guesser? Explain.

**Practice Problem 5.7A**

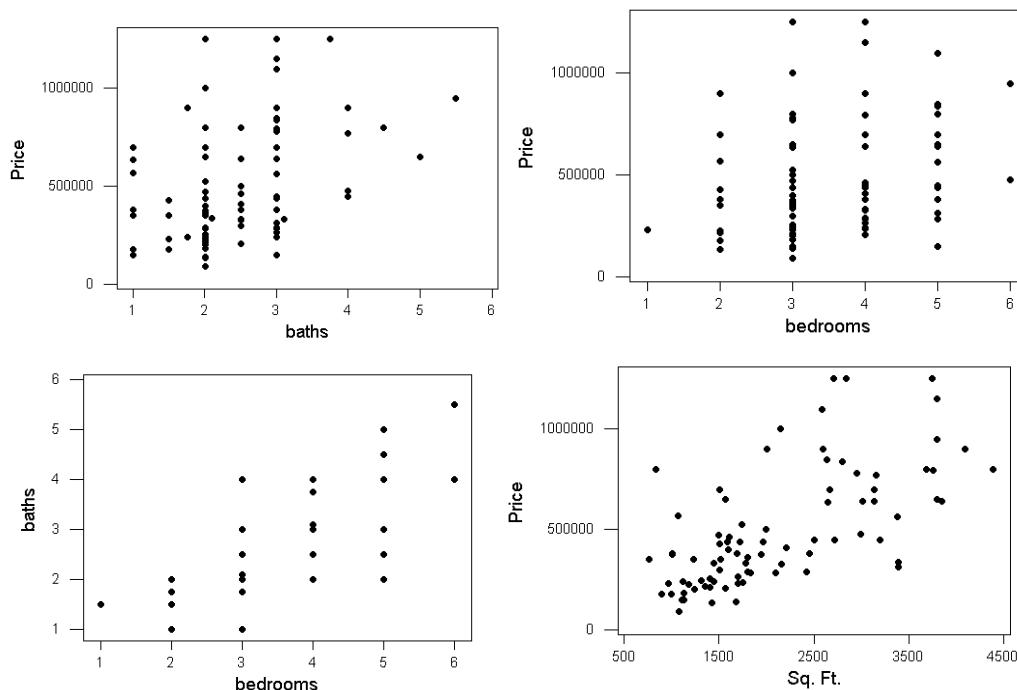
Suppose that we record the midterm exam score and the final exam score for every student in a class. What would the value of the correlation coefficient be if every student in the class scored:

- Ten points higher on the final than on the midterm?
- Five points lower on the final than on the midterm?
- Twice as many points on the final exam as on the midterm?

Explain your answer in each part. [Hint: You might first want to draw yourself a scatterplot of hypothetical data that fit the stated conditions.]

**Practice Problem 5.7B**

(a) The following scatterplots look at the relationships between house prices and four other variables. How does the strength of the linear relationship between price and square footage compare to the strength of the relationships in the first 3 graphs?



- (b) The correlations for these four graphs are  
0.284, 0.394, 0.649, 0.760

Which correlation coefficient do you think corresponds to which graph? Explain your reasoning.  
(Note: Each graph has the same number of houses, but you may have multiple houses indicated by an individual dot.)

### Investigation 5.8: Height and Foot Size

Criminal investigators often need to predict unobserved characteristics of individuals from observed characteristics. For example, if a footprint is left at the scene of a crime, how accurately can we estimate that person's height based on the length of the footprint? To investigate this possible relationship, data were collected on a sample of students in an introductory statistics class.

- (a) Identify the observational units, explanatory variable, and response variable in this study.

Observational units:

Explanatory variable:

Response variable:

Below are the heights (in inches) of 20 students in a statistics class:

74 66 77 67 56 65 64 70 62 67 66 64 69 73 74 70 65 72 71 63

#### Predicting Heights

- (b) If you were trying to predict the height of a statistics student based on these observations, what value would you report?

- (c) Using the method in (b), would you always predict a statistics student's height correctly?

**Definition:** A residual is the difference between the predicted value and the observed value. If we let  $y_i$  represent the  $i^{\text{th}}$  observed value and  $\hat{y}_i$  represent the predicted or “fitted” value, then

$$\text{residual}_i = y_i - \hat{y}_i$$

- (d) The table below shows the residuals for each of the above heights if we use the mean height, 67.75 inches, as the predicted value for each person. Calculate the last two residuals to complete the table.

| Height   | 74    | 66    | 77   | 67   | 56     | 65    | 64    | 70   | 62    | 67   |
|----------|-------|-------|------|------|--------|-------|-------|------|-------|------|
| Residual | 6.25  | -1.75 | 9.25 | -.75 | -11.75 | -2.75 | -3.75 | 2.25 | -5.75 | -.75 |
| Height   | 66    | 64    | 69   | 73   | 74     | 70    | 65    | 72   | 71    | 63   |
| Residual | -1.75 | -3.75 | 1.25 | 5.25 | 6.25   | 2.25  | -2.75 | 4.25 |       |      |

How often did we overestimate? How often did we underestimate?

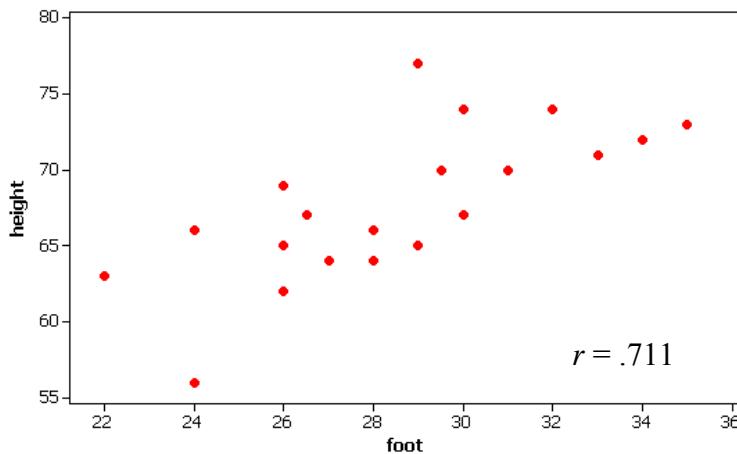
- (e) How does the observed value compare to the predicted value when the residual is positive? How about when the residual is negative?

- (f) How might you combine the residuals to measure the overall prediction error for these 20 students?
- (g) Calculate the sum of the residuals from (d). Explain why taking the sum of residuals is not a useful way to measure overall prediction error.
- (h) Show mathematically that the sum of residual from the sample mean for *any* dataset equals zero. In other words, show that  $\sum_{i=1}^n (y_i - \bar{y}) = 0$ .
- (i) Suggest two ways to get around the problem revealed in (g) and (h). In other words, suggest something related to but different from summing residuals to use as a useful measure of overall prediction error.
- (j) Suppose that you want to use a single number (call it  $m$ ) for predicting height. Use calculus to determine  $m$ , as a function of the data  $y_i$ 's, to minimize the sum of squared residuals from that prediction. [Hints: You are choosing  $m$  to minimize  $S = \sum_{i=1}^n (y_i - m)^2$ . Take the derivative with respect to  $m$ , set it equal to zero, and solve for  $m$ .] Interpret your answer for  $m$ .

The Excel Exploration after this investigation asks you to consider minimizing other criteria, such as the sum of absolute residuals.

### Predicting Heights from Footlengths

To see whether we can make better predictions of height by taking foot length into account, consider the following scatterplot of the height (in inches) and foot length (in centimeters) for the sample of 20 statistics students.



- (k) Describe the association between height and foot length exhibited in this scatterplot. Is the association what you would have expected? Explain.

Open the [Analyzing Two Quantitative Variables](#) applet to see the scatterplot of the 20 students' height and foot measurements. Check the **Show Movable Line** box to add a blue line to the scatterplot. The

equation for this initial line,  $\hat{\text{height}} = 67.75 + 0 \text{ foot size}$ , predicts the same height for all 20 students as you did in (d). (Note: the “hat” over height indicates that the equation gives values for *predicted* height.)

### Analyzing Two Qu:

Sample data:  
(explanatory, response)

| footlength | height |
|------------|--------|
| 32.0       | 74     |
| 24.0       | 66     |
| 29.0       | 77     |
| 30.0       | 67     |
| 24.0       | 56     |
| 26.0       | 65     |
| 27.0       | 64     |
| 29.5       | 70     |
| 26.0       | 62     |

Use My Data | Clear

If you now place your mouse over the green square on one of the ends of the line and drag, you can change the slope of the line. You can also use the mouse to move the green square in the middle of the line up and down vertically to change the intercept of the line.

- (l) Move the line until you feel your line “best” summarizes the relationship between height and foot length for these data. Write down the final equation for your line. [Hint: Use good statistical notation, which means to use variables names (not generic  $x$  and  $y$ ) and to put a “hat” over the response variable to indicate prediction.]

- (m) Did everyone in your class obtain the same line/equation?

(n) Does your line provide a better fit than your neighbor's? Suggest a criterion for deciding which line "best" summarizes the relationship.

(o) Check the **Show Residuals** box to visually represent these residuals for your line on the scatterplot. The applet also reports the sum of the absolute residuals (or SAE, the sum of absolute errors). Record this SAE value for your line. What is the best SAE in the class? (Does "best" correspond to the smallest or the largest value of SAE?)

(p) A more common criterion for determining the "best" line is to instead look at the sum of the *squared* residuals (or sum of squared errors, SSE). Check the **Show Squared Residuals** box to visually represent them and to determine the SSE value for your line. What is the best SSE in the class?

(q) Continue to adjust your line until you think you have minimized the sum of the squared residuals. Report your new equation and new SSE value.

(r) Now check the **Show Regression Line** box to determine and display the equation for the line that actually does produce the smallest possible sum of squared residuals. Report its equation and SSE value. Did everyone obtain the same line/equation this time? How does it compare to your line? (You can also display the residuals and the squared residuals for this line.)

Equation:

SSE:

Comparisons:

**Definition:** The line that minimizes the sum of squared residuals is called the [least squares line](#), or simply the [regression line](#), or even the *least squares regression line*.

(s) Suggest a technique for determining (based on the observed data  $x_i$ 's and  $y_i$ 's) the values of the slope and the intercept that minimize the SSE.

**Least Squares Regression Line: Derivation of Coefficients**

The [least squares line](#)  $\hat{y} = b_0 + b_1x$  is determined by finding the values of the coefficients  $b_0$  and  $b_1$  that minimize the sum of the squared residuals,  $SSE = \sum(y_i - \hat{y}_i)^2 = \sum_{i=1}^n(y_i - b_0 - b_1x_i)^2$ .

(t) Take the derivative with respect to  $b_0$  of the expression on the right. Then take the derivative of the original expression with respect to  $b_1$ . [Hints: Use the chain rule, and remember to treat the data values  $x_i$ 's and  $y_i$ 's as constants.]

(u) Set these (partial) derivatives equal to zero and solve simultaneously for the values of  $b_0$  and  $b_1$ .  
[Hints: Solve the first equation for  $b_0$ . Then solve the second equation for  $b_1$  (substituting in the expression for  $b_0$  using the summation notation).]

You should have found expressions equivalent to the following.

$$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} \quad \text{and} \quad b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

With a little bit more algebra, you can show that the formulas for the least squares estimates of the intercept coefficient  $b_0$  and slope coefficient  $b_1$  simplify to:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = r s_y / s_x$$

(v) For the sample data on students' foot lengths ( $x$ ) and heights ( $y$ ), we can calculate these summary statistics:  $\bar{x} = 28.5$  cm,  $s_x = 3.45$  cm,  $\bar{y} = 67.75$  in and  $s_y = 5.00$  in, with  $r = 0.711$ . Use these statistics and the formulas above to calculate the coefficients of the least-squares regression line. Confirm that these agree with what the applet reported for the equation of the least squares line.

(w) Use this least-squares regression line to predict the height of a person with a 28 cm foot length. Then repeat for a person with a 29-cm foot length. Calculate the difference in these two height predictions. Does this value look familiar? Explain.

(x) Provide an interpretation of the slope coefficient ( $b_1$ ) in this context.

(y) Provide an interpretation of the intercept coefficient ( $b_0$ ) in this context. Is such an interpretation meaningful for these data? Explain.

**Discussion:** The slope coefficient of 1.03 indicates that the predicted height of a person increases by 1.03 inches for each additional centimeter of foot length. In other words, if one person's foot length is one cm longer than another's, we predict this person to be 1.03 inches taller than the other person. Notice we are being careful to talk about the "predicted" change or "average" change, because these are estimates based on the least squares line, not an exact mathematical relationship between height and foot length. The intercept coefficient can be interpreted as the predicted height for a person whose foot length is zero – not a very sensible prediction in this context. In fact, the intercept will often be too far outside the range of  $x$  values for us to seriously consider its interpretation.

- (z) Use the least squares regression line to predict the height of someone whose foot length is 44 cm. Explain why you should not be as comfortable making this prediction as the ones in (w).

**Definition:** [Extrapolation](#) refers to making predictions outside the range of the explanatory variable values in the data set used to determine the regression line. It is generally ill-advised.

### Evaluating the model

One way to assess the usefulness of this least-squares line is to measure the improvement in our predictions by using the least-squares line instead of the  $\bar{y}$  line that assumes no knowledge about the explanatory variable.

- (aa) In the applet, uncheck and recheck the **Show Movable Line** box. Check the **Show Squared Residuals** box to determine the SSE if we were to use  $\bar{y}$  as our predicted value for every  $x$  (foot size). Record this value.

$$SSE(\bar{y}) =$$

- (bb) Recall the SSE value for the regression line. Determine the *percentage change* in the SSE between the  $\bar{y}$  line and the regression line:

$$100\% \times [(SSE(\bar{y}) - SSE(\text{least-squares})) / SSE(\bar{y})] =$$

**Definition:** The preceding expression indicates the reduction in the prediction errors from using the least squares line instead of the  $\bar{y}$  line. This is referred to as the [coefficient of determination](#), interpreted as the percentage of the variability in the response variable that is explained by the least-squares regression line with the explanatory variable. This provides us with a measure of how accurate our predictions will be and is most useful for comparing different models (e.g., different choices of explanatory variable). The coefficient of determination is equal to the square of the correlation coefficient and so is denoted by  $r^2$  or  $R^2$  (and is often pronounced "r-squared.")

Another measure of the quality of the fit is  $s$ , the standard deviation of the residuals. This is a measure of the unexplained variability *about the regression line* and gives us an idea of how accurate our predictions should be (the actual response should be within  $2s$  of the predicted response). If  $s$  is much smaller than the variability in the response variable ( $s_y$ ) then we have explained a good amount of variability in  $y$ . Most statistical packages report  $s$ , or it can be found from  $\sqrt{SSE/(n - 2)}$ .

- (cc) Determine and interpret the value of  $s$  for these data. [What are the units?]

### Study Conclusions

There is a fairly strong positive linear association between the foot length of statistics students and their heights ( $r = 0.711$ ). To predict heights from foot lengths, the least-squares regression line is

$\hat{height} = 38.3 + 1.03 \text{ foot}$ . This indicates that if one person's foot length measurement is one centimeter longer than another, we will predict that person's height to be 1.03 inches taller. This regression line has a coefficient of determination of 50.6%, indicating that 50.6% of the variability in heights is explained by this least squares regression line with foot length. The other 49.4% of the variability in heights is explained by other factors (perhaps including gender) and also by natural variation. So although the foot lengths are informative, they will not allow us to perfectly predict the heights of the students in this sample. The value of  $s$  is 3.61 inches, meaning we should be able to predict a person's height within 3.61 inches based only on the size of his or her foot.

### Technology Detour – Determining Least Squares Regression Lines

#### In R

To calculate intercept and slope: > `lm(response~explanatory)`

To then superimpose regression line on scatterplot: > `abline(lm(response~explanatory))`

### Practice Problem 5.8

For the Cat Jumping data set from Investigation 5.6:

- Calculate and interpret the correlation coefficient between velocity and body mass.
- Square the correlation coefficient to obtain  $r^2$ . Interpret the coefficient of determination in context.
- Determine and interpret the slope of the least squares regression line.
- Determine and interpret the intercept of the least squares regression line. Explain what this value might signify in this context.
- Determine and interpret the value of  $s$ .

## Applet Exploration: Behavior of Regression Lines

Open the [Analyzing Two Quantitative Variables](#) applet. Recall the 20 observations are height and foot length for a sample of statistics students.

- (a) Check the **Show Regression Line** box. Does the regression line appear to do a reasonable job of summarizing the overall linear relationship in these data observations?
- (b) If we were to add an observation with height 60 in and foot length 35 cm, predict whether, and if so how, the regression line will change.
- (c) Enter these coordinates in the Add/remove observations box ( $x = 35$ ,  $y = 60$ ) and press the **Add** button. Report the new regression equation.
- (d) Check the **Move observations** box. Click on this new observation (it should change color) and hold the mouse button down and move the mouse vertically in both directions to change the  $y$  value of the observation (try hard not to change the  $x$  value). The applet automatically recalculates the new regression line depending on the new location of the point. Is it possible to make the regression line have a negative slope? Does the regression line appear to be affected by the location of this point? Is the impact strong or weak? Does this match your prediction?
- (e) Delete this new point and/or press the **Revert** button to return to the original data set. Now focus on the point located at (29, 65). If we move this point vertically, predict how the regression line will change. Do you think the change will be as dramatic as in (c)?
- (f) Repeat (d) using the point located at (29, 65). Does the regression line appear to be affected by the location of this point? Is the impact strong or weak (especially compared to the impact you witnessed in (d))?
- (g) Which point was more *influential* on the equation of the regression line: (35, 60) or (29, 65)? Suggest an explanation for why the point you identified is more influential, keeping in mind the “least-squares criterion.”

## Excel Exploration: Minimization Criteria

Another way to get around the problem revealed in (g) and (h) of Investigation 5.8, that the sum of residuals from the mean always equals zero, is to examine the sum of the *absolute* deviations.

- (a) Using your results in the table in question (d) of Investigation 5.8, calculate the sum of the absolute values of the residuals, using the mean height as the predicted value.

Let's investigate whether there is a better number (call it  $m$ ) than the mean to use for predicting height, in order to minimize our prediction errors. In general, this sum of absolute residuals (or SAE, sum of absolute errors) can be written:

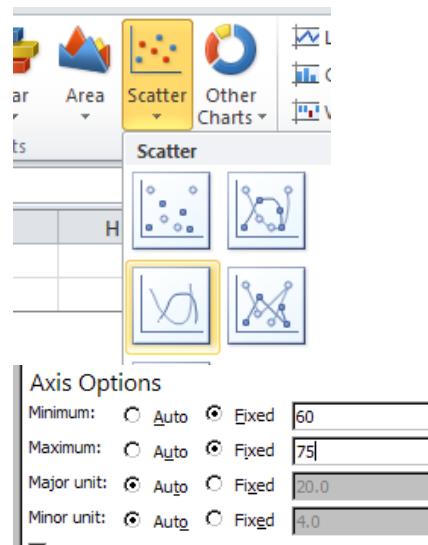
$$SAE = \sum_{i=1}^n |y_i - m|$$

The problem with using the SAE is that it is not easily differentiable, so we cannot use calculus to determine the optimal value for  $m$ . Instead, we will use Excel to explore the behavior of SAE as a function of  $m$ . We will look for the value of  $m$  that minimizes SAE.

- (b) Open the Excel spreadsheet document [Heights.xls](#). Notice that the 20 heights are in column A (sorted), and column B contains candidates for  $m$  from 65 to 70, in increments of 0.01. Click on cell C2, and notice that it contains a formula for  $SAE(m)$ , which refers to the data values in column A and to the value of  $m$  from B2 (namely,  $m = 65$  in that cell). Use Excel's "fill down" feature to calculate the values of  $SAE(m)$  for the remainder of column C. [Hint: You can do this in one of several ways: With the C2 cell selected, double click on the box in the lower right corner, or pull the right corner of the highlighted C2 cell down to the end of the column, or highlight C2 and the cells to be filled and choose Edit> Fill> Down.] Confirm that the SAE value for the sample mean (67.75) agrees with what you calculated in (a).

- (c) To produce a graph of SAE as a function of  $m$ : Highlight columns B and C. **Choose Insert > Scatter** and then choose the first option in the second row (Scatter with Smooth Lines) to construct a graph of this function.

[Hint: After you create the graph, you may want to change the  $x$ -min and  $x$ -max values by double clicking on the horizontal axis and changing the Minimum and Maximum under Axis Options.]



Reproduce a rough sketch of this graph below, and comment on the shape of this function. Does it seem to follow a familiar form, such as linear or parabolic or exponential? Explain.

(d) Does there seem to be a unique value of  $m$  that minimizes this SAE( $m$ ) function? If so, identify it. If not, describe all values of  $m$  that minimize the function. Also report the value of the SAE for this optimal value of  $m$ . [Hint: Search through the entries in column C to find the smallest value, and then read off the corresponding value in column B.]

Optimal  $m$  =

SAE at the optimal  $m$  =

(e) Look through the sorted heights in column A. What do you notice about where this optimal value of  $m$  is located in the list of sorted heights? Does this remind you of a familiar statistic?

(f) Suppose the tenth largest height (in row 11) had been 66 inches instead of 67. Make this change to the spreadsheet (cell A11), and note that the graph of the SAE function is updated automatically. Now is there a unique value of  $m$  that minimizes this SAE( $m$ ) function? If so, identify it. If not, describe all values of  $m$  that minimize the function, and also report the (optimal, minimum) value of the SAE.

Optimal  $m$  =

SAE at the optimal  $m$  =

(g) Return the tenth largest height back to 67 inches. Now suppose the tallest person in the sample had been 80 inches tall instead of 77. Make this change to the spreadsheet (cell A21) and note how the graph of the SAE function changes. Also report the value of the SAE.

Optimal  $m$  =

SAE at the optimal  $m$  =

(h) Based on these findings, make a conjecture as to how one can determine, from a generic set of 20 data values, the value that minimizes the sum of the absolute prediction errors from the data values.

(i) Consider a new minimization criterion: mean absolute error (MAE) criterion, which divides the SAE value by the number of observations. How would the MAE( $m$ ) function compare to the SAE( $m$ ) function? Would its minimum be achieved at the same value(s) of  $m$ ? Explain.

**Discussion:** You should have found that the SAE function is *piecewise linear*. In other words, the function itself is not linear, but its component pieces are linear. In this case the function changes its slope precisely at the data values. The SAE function (and MAE function) are minimized at the *median* of the data and also at any value between (and including) the two middle values in the dataset (the 10<sup>th</sup> and 11<sup>th</sup> values in this case with  $n = 20$  values). It is a convenient convention to define the median as the average (midpoint) of those two middle values. The MAE has a bit more natural interpretation – the average deviation of the values in the data set from the value of  $m$ .

Another criterion is to find the value of  $m$  that minimizes the *sum of the squared deviations*.

(j) Recall from (j) of Investigation 5.8 the value of  $m$  that minimizes the sum of the squared prediction errors. Report this value of  $m$  for the height data.

(k) Make sure that the data values in column A are back to their original values. Then click on cell D2 in the Excel spreadsheet, and notice that it gives the formula for the SSE (sum of squared errors) function. Fill this formula down the column and create a second graph to display the behavior of this function. [Hints: Click on column B and then hold the mouse down when you click on column D to highlight just those two columns before you insert the graph. You will probably want to double click on the y-axis scale to change its minimum value.] Does this function have a recognizable form, such as a polynomial? What is its shape? At what value of  $m$  is the function minimized? Does this value look familiar?

(l) Suppose there had been a clerical error in entering the height of the tallest student. Change the height of the tallest student from 77 inches to 770 and investigate the effect on the SSE function. Has the value of  $m$  that minimizes SSE changed? By how much? How does this compare to the effect of this clerical error on the SAE function and where it is minimized?

**Discussion:** Although the SAE criteria would be more resistant to outliers, it does not always lead to a unique minimum (e.g., an even number of data values where the values of the middle pair are not equal). This has led to the more common criterion of *least squares*.

See the exercises for investigating additional minimization criteria, such as the median of absolute errors and maximum of absolute errors.

### Investigation 5.9: Money-Making Movies

Is there a tendency for movies that garner better reviews to also earn more money at the box office? *USA Today* (Wloszczyna & DeBarros, Feb. 25, 2004) investigated this by determining a rating score for movies released in 2003 based on a compilation of movie reviews published in 20 major newspapers and magazines for over 300 movies. The [movies03.txt](#) data file contains these scores and how much money the movie made at the box office, in millions of dollars. A high composite score indicates that most critics loved the movie, and a low score indicates that most critics panned the movie.

- (a) Identify the observational units in this study. Also identify the explanatory and response variables, and classify them as categorical or quantitative.

Observational units:

Explanatory:

Response:

- (b) Produce a scatterplot to determine whether the critic scores appear to help predict the box office gross. Describe the relationship between the two variables as exhibited in the scatterplot.

- (c) Identify the two or three points that you believe have the largest (in absolute value) residuals. Identify these movies by name

- **In R:** You can use a command like `> movies03[score > 90 & box.office > 250, ]`

What does it mean for these movies to have such large residuals?

- (d) Determine, report, and interpret the value of the correlation coefficient.

- (e) Use technology to determine the least squares regression line for predicting the box office gross from the composite critics' score. Report the equation for this line. Also interpret the slope and intercept of this line in context.

Equation of line:

Interpretation of slope:

Interpretation of intercept:

- (f) Report the value of  $r^2$  and  $s$  provide interpretations in this context.

### Technology Detour – Creating a Coded Scatterplot with Separate Lines

#### In R

- Create the coded scatterplot: > `plot(box.office~score, col=rating)`
  - Add the separate lines for each category, e.g.,:  
`> abline(lm(box.office[rating=="G"]~score[rating=="G"]))`  
To cycle through the categories, create a column for the four categories and create a loop:  
`> code=c("G", "PG", "PG-13", "R")  
> for (i in 1:4) {  
 abline(lm(box.office[rating==code[i]]~score[rating==code[i]]),  
 col=i) }`
- To add a legend: > `legend("topleft", legend=code, col=1:4, pch=1)`

- (g) Describe what this coded scatterplot reveals about whether the relationship between box office income and critic scores differs across the various rating categories. In particular, does any rating category tend to have higher box office values than you would expect for the score they received from the critics? Is there a rating category tending toward lower box office revenues than expected? Explain.

- (h) Repeat (g) using the movie genre (e.g., comedy) as the categorical variable.

(i) In analyzing these movies, the researchers also looked at the data set after removing the 6 movies that earned more than \$200 million. Subset the data (e.g., Investigation 2.1) and then examine the scatterplot, correlation coefficient, and least-squares regression line for predicting the movie revenue from the critics' rating score for these data.

New correlation coefficient:

New least-squares regression line:

(j) Describe the effect of removing these 6 movies from the analysis.

**Definition:** An observation or set of observations is considered influential if removing the observation from the data set substantially changes the values of the correlation coefficient and/or the least squares regression equation. Typically, observations that have extreme explanatory variable outcomes (far below or far above  $\bar{x}$ ) are potentially influential. To measure the influence of an observation, it is removed, and measures are calculated for how much the summary results change. It is *not* always the case that the points with the largest residuals are the most influential.

In this example you should have seen that removing those six movies actually makes the relationship between box office income and critic scores much weaker ( $r$  drops from 0.42 to about 0.3).

### Study Conclusions

There does appear to be a weak relationship between the composite critics' scores and the amount of money the movie makes at the box office, with higher rated movies making more money. If the composite critics' score is 10 points higher, we predict the movie will make about 18.57 million more dollars. It is interesting to note that this regression line will tend to overestimate the amount of revenue for an R rated movie and underestimate the revenue of action movies. If the top 6 grossing movies of 2003 (*Bruce Almighty*, *Finding Nemo*, *Pirates of the Caribbean*, *The Lord of the Rings III*, *The Matrix Reloaded*, and *X2: X-Men United*) are removed, the relationship is not as strong, but still shows a weak positive linear association ( $r = 0.3$ ).

### Practice Problem 5.9

- Which do you think will be more resistant to outliers, the regression line that minimizes the sum of squared errors or the regression line that minimizes the sum of the absolute errors? Explain.
- Investigate your conjecture using the [Analyzing Two Quantitative Variables](#) applet. Write a short paragraph summarizing your analysis and your observations.

## Section 5.3 Summary

In this section you considered the association between two quantitative variables. Because this was your first encounter with this type of analysis scenario, we returned to several themes from Chapter 1, including starting with graphical displays and numerical summaries. You learned that the relevant graphical display for describing the association between two quantitative variables is a *scatterplot*, and that some features to look for in a scatterplot are the *direction*, *strength*, and *linearity* of the association. The most common numerical summary of a linear association is the *correlation coefficient*, and you studied some properties of this measure. Then you turned your attention to fitting a linear model to the data, useful for predicting the value of the response variable based on an explanatory variable outcome. You studied the *least squares criterion* for determining this *regression* line and derived the equations for estimating the coefficients from sample data. With the regression line in hand, you can make predictions for the response variable outcome based on the explanatory variable value. Just keep in mind that these predictions should be made only for  $x$  values within the range of the original data (and not *extrapolating* beyond the original explanatory variable values. Finally, you examined properties of regression lines, such as the *coefficient of determination*, which reveals how much of the variability in the response variable is explained by the regression line, and *influential observations*, which have substantial effect on the regression line and typically arise from observations with extreme explanatory variable values.

## SECTION 4: INFERENCE FOR REGRESSION

In the previous section you learned graphical and numerical techniques for describing the relationship between two quantitative variables, namely scatterplots and correlation coefficients. You also learned a method for fitting a linear model to predict the value of one quantitative variable from another. But you only *described* sample data in the previous section. In this section, you will again make the transition to *inference*: what can we conclude about a larger population or about the statistical significance of the association based on the sample results?

### Investigation 5.10: Running Out of Time

Many communities offer racing events for amateur runners, such as 5K (approximately 3.1 miles) races. At the end of the event, awards are often given out to runners with the fastest times, often to the top three finishers in different age groups. *But does running time really change with age for amateur runners?* To investigate this, one of the authors downloaded the race results for a local 5K race in May 2013 ([Talley5K2013.txt](#)). Data for 248 runners included age (in years) and finish time (in minutes).

- (a) Identify the explanatory and response variables for this research question.

EV:

RV:

- (b) Do you expect the relationship between finish time and age to be positive or negative? Do you expect it to be linear (which implies what)? Do you think the association between these variables will be statistically significant?

Direction:

Form:

Significance:

- (c) Use technology to produce a scatterplot of *finish time* vs. *age*. [You can copy the data into Excel, using Data > Text to Columns and then delimiting by spaces, and then copy the last two columns (age, time in that order) into the [Analyzing Two Quantitative](#) variables applet.] Do you want to change any of your answers to (b) in light of this graph? Also discuss any other unusual features of the data.

- (d) Remove the outlier (the name was a duplicate and the time was later dropped from the official published results), and determine the least-squares regression equation for predicting finish time from age. Interpret the slope coefficient in context.

(e) Suppose we consider the remaining 247 runners to be a random sample from some larger population. Is it possible that there is no association between age and finishing time in this larger population, but we just happened to find such an association in this sample? How might we investigate how plausible this explanation is?

### Terminology Detour

To differentiate between the population regression line and the sample regression line, we will continue to use  $b_0$  and  $b_1$  to denote the sample statistics for the intercept and slope, respectively (or alternatively  $\hat{\beta}_0$  and  $\hat{\beta}_1$ ). The corresponding population parameters will be denoted by  $\beta_0$  and  $\beta_1$ .

(f) If there were no relationship between height and age in the population, what would this imply about the value of the population slope? State null and alternative hypotheses to reflect the research question that 5K runner finishing times change with age. (Do you think they will increase or decrease?)

We now want to access the statistical significance of the association between variables in this sample. Following our usual logic, we will approach this question by assessing how unusual it would be to obtain a sample slope coefficient of  $b_1 = 0.140$  (or more extreme) with a random sample of 247 runners from a population with no association between the variables (i.e., with a population slope coefficient of  $\beta_1 = 0$ ). We will do this by creating such a population and repeatedly taking random samples from it, calculating the sample slope each time.

First, let's think about what we want this population to look like.

(g) Determine the sample mean and standard deviation of the remaining individual variables in our sample:

|                    |        |                      |
|--------------------|--------|----------------------|
| <i>Finish time</i> | Mean = | Standard deviation = |
| <i>Age</i>         | Mean = | Standard deviation = |

Also determine the value of  $s$  from the regression model ( $\sqrt{SSE/(n - 2)}$ ) in (d) = \_\_\_\_\_

So one approach is to create a population that has similar characteristics as this sample, but with a population slope coefficient of zero.

*Technical note:* You also want think about whether you want to model random sampling from a bivariate population, or random sampling of finishing times at each age. We will start with the first approach, but classical regression theory (Investigation 5.12) assumes the second.

In the [Analyzing Two Quantitative Variables](#) applet, check the **Create Population** box in the bottom left corner. We will create a population of runners where there is no relationship between *finish time* and *age*. To match the above context:

- Set the population **intercept** to be equal to the mean of the *time* variable, **32.36**. Because we are forcing  $\beta_1 = 0$ , this says that the population regression line is constant at  $\bar{y} = 32.36$  minutes.
- Set the **x mean** and **x standard deviation** to match the values you found in (g) for *age* after removing the outlier.
- Set the value of **sigma**, representing the variability about the population regression line, to equal the value of *s* from the regression model (should be smaller than the SD(Y) you found in (g)).
- Press the **Create Population** button.
- Press **Rescale**.

So the population of 20,000 runners we are creating here matches the characteristics of the observed sample data. The key difference here is that we are assuming the population slope is equal to zero.

(h) Check the **Show Sampling Options** box and set the sample size to be 247 and press the **Draw Samples** button. Did you obtain the same sample regression line (in blue) as from the observed race results or as in the hypothesized population (in red)?

(i) Press the **Draw Samples** button again. Make sure **Show Regression Line** is checked and record the regression line displayed in blue. Did you obtain the same sample regression line as in (h)?

**Discussion:** These questions reveal once again the omnipresent phenomenon of *sampling variability*. Just as you have seen that the value of a sample mean varies from sample to sample, the value of a sample proportion varies from sample to sample, and the value of a chi-square statistic varies from sample to sample, now you see that the value of a sample regression line also varies from sample to sample. Once again we can use simulation to explore the long-run pattern of this variation, approximating the *sampling distributions* of the sample slope coefficients and the sample intercept coefficients.

(j) Change the **Number of Samples** from 1 to 20. Press the **Draw Samples** button. Describe the pattern of variation that you see in the simulated regression lines.

(k) Now change the **number of samples** to **1000** and press the **Draw Samples** button. On the next page, describe the shape, mean, and standard deviation of this distribution.

### Analyzing Two Quantitative Variables

#### Population inputs

|                       |        |
|-----------------------|--------|
| population slope:     | 0      |
| population intercept: | 32.364 |
| population x mean:    | 38.506 |
| population x std:     | 16.849 |
| population sigma:     | 8.265  |

**Create Population**

Note: These values should appear in the applet when you check Create Population if you had pasted in the sample data first.

Shape:

Mean:

Standard deviation:

(l) Is the *mean* of this distribution of sample slopes roughly what you expected? Explain.

(m) Now change the value of **sigma** from 8.265 to **4.29** and press **Create Population**. How does this change the scatterplot?

(n) How do you think this will change the behavior of the distribution of sample slopes? (shape, center, variability)

(o) Press the **Draw Samples** button. When you scroll down you should see the original dotplot at the bottom and a new dotplot for this value of  $\sigma$  at the top. Was your conjecture in (m) correct? Be sure to comment on shape, center, and variability.

(p) Change the value of **sigma** back to **8.265** (and press **Create Population**) but change the value of **x std** from 16.89 to **8.45**. Press the **Create Population** button. How does this change the scatterplot? How do you think this will change the behavior of the distribution of sample slopes and the distribution of the sample intercepts?

Change to scatterplot:

Prediction:

(q) Press the **Draw Samples** button. Was your conjecture in (p) correct?

(r) Change the value of **x std** back to 16.89 (**Create Population**) and change the **sample size** from 247 to **125**. Conjecture what will happen to the sampling distribution of the sample slopes.

(s) Press **Draw Samples**. Was your prediction correct?

(t) Summarize how each of these quantities affect the sampling distribution of the sample slope:

- The variability about the regression line,  $\sigma$  (sigma)
- The variability in the explanatory variable,  $SD(X)$
- The sample size,  $n$ .

You should have made the following observations (when the technical conditions are met):

- The distribution of sample slopes is approximately normal.
- The mean of the distributions of sample slopes equals the population slope  $\beta_1$ .
- The variability in the sample slopes increases if we increase  $\sigma$ .
- The variability in the sample slopes increases if we decrease  $s_x$ .
- The variability in the sample slopes decreases if we increase  $n$ .

(u) Explain why each of the last 3 observations make intuitive sense. [You may want to recall the pictures from the Applet Exploration.]

- (v) Are the last three observations consistent with the following formula for the standard deviation of the sample slope,  $SD(b_1)$ ? Explain.

$$SD(b_1) = \sigma \sqrt{\frac{1}{(n-1)s_x^2}}$$

Back to the question at hand: Is it plausible that the observed sample slope ( $b_1 = 0.140$ ) came from a population with no association between the variables and so a population slope  $\beta_1 = 0$ ?

- (w) Return to (or recreate) the dotplot of the sample slopes for the first simulation. Where does 0.140 fall in this null distribution? Is it plausible that the population slope is really zero and we obtained a sample slope as big as 0.140 just by chance? How often did such a sample slope occur in your 1000 samples? In all the samples obtained by the class?

- (x) Change the **population slope** to **.10** and press **Create Population**. Draw **1000** samples and examine the new hypothesized population and the new sampling distribution of the sample slopes. (You may want to press **Rescale**.) Where is the center? Roughly how often did you get a sample slope as big as 0.140 or bigger? Does it seem plausible that the students' regression line came from a population with  $\beta_1 = 0.10$ ?

### Study Conclusions

These sample data provide extremely strong evidence of a relationship between age and finishing time for the population of amateur runners similar to those in this sample. The above simulation shows that if there were no relationship in the population ( $\beta_1 = 0$ ), then we would pretty much never see a sample slope as large as 0.140 just by chance (random sampling). We should be a bit cautious in generalizing these results to a larger sample as they were not a random sample. We might want to limit ourselves to 5K racers in similar types of towns.

**Discussion:** As you have seen several times in this course, the sampling distribution of a statistic can often be well described by a mathematical model. This is also true in the regression setting. As when we focused on analyzing the individual population mean or comparing two population means, we need to first consider a way to estimate the (nuisance) standard deviation parameter.

- (y) If  $\sigma$  represents the common standard deviation of the response values about the regression line, suggest a way of estimating  $\sigma$  from the sample data.

**Discussion:** If the conditions for the basic regression model (Investigation 5.12) are met, then if we standardize the sample slopes:

$$t_0 = \frac{b_1 - \beta_1}{SE(b_1)}$$

where  $SE(b_1) = s \sqrt{\frac{1}{(n-1)s_x^2}}$  and  $s = \sqrt{\sum_{i=1}^n \frac{(residuals_i)^2}{n-2}}$ ,

these standardized values can be shown to be well approximated by a  $t$ -distribution with  $n - 2$  degrees of freedom. Note that we lose a second degree of freedom because we are estimating both the slope and the intercept from the sample data.

- (z) Using the standard deviation for the slopes you found in (k), calculate the test statistic.

Once we have the  $t$ -statistic or  $t$ -ratio, we can use the Student's  $t$ -distribution to compute one- or two-sided p-values depending on the research question regarding the slope.

*Note:* The  $t$ -distribution with  $n - 2$  degrees of freedom also provides a reasonable model for a randomization distribution that repeatedly re-randomizes the values of the explanatory variable and looks at the distribution of the resulting sample slopes (see Applet Exploration).

- (aa) Confirm the  $t$ -ratio and p-value calculations using technology.

- In R: Use > `summary(lm(time~age))`

### Practice Problem 5.10A

Recreate this analysis (simulation as in (h) and (w) and  $t$ -test as in (z)) without first removing the outlier. Do the results change much? If so, how? In the expected manner? Explain.

### Practice Problem 5.10B

The official published results for the 2013 race can be found [here](#).

- Load these data into your technology and then compare the two datasets.
- Identify a second difference (apart from Scharlach\_1 disappearing) between the unofficial and the official results.
- How does the strength of evidence of a relationship between *time* and *age* differ for this dataset?

**Investigation 5.11: Running Out of Time (cont.)**

In the previous investigation, we analyzed the statistical significance of our sample data by replicating how often we might get a *random sample* of 247 runners (after removing an extreme outlier) with a slope this extreme from a larger population of runners that did not have an association between *finishing time* and *age*. In other words, we assumed there was no association between *finishing time* and *age* and the pairing of these values for our runners was just arbitrary. This suggests another way of assessing the statistical significance of our results.

- (a) Paste the [Talley5K2013.txt](#) data file into the [Analyzing Two Quantitative Variables](#) applet. Press **Use Data** and make sure *age* is used as the explanatory variable (if not press the **(response, explanatory)** button to toggle their roles). Remove the outlier and check the **Show Regression Line** box and confirm the equation.

Another way to investigate whether the association observed in these sample data could have arisen by chance is a randomization test that shuffles the outcomes for one of the variables and reexamines the correlation coefficient and/or sample slope for the shuffled or re-randomized data.

- (b) Check the **Show Shuffle Options** box. Press the **Shuffle Y-values** box. Select the **Plot** radio button. Describe the blue line that appears. Is the association as strong as the one that we found in the actual data? How are you deciding?

- (c) Press the **Shuffle Y-values** button 5-10 more times. Do you ever find a negative association? Are any of the associations as strong as the one observed by the student group? What pattern is beginning to emerge in the blue regression lines on the scatterplot?

- (d) Change **Number of Shuffles** to some large number (like 1000 or the difference between how many you have done so far and 1000) and press the **Shuffle Y-values** button. Describe the shape, center, and variability of the resulting distribution of re-randomized slopes. How does this distribution compare to the one you found in Investigation 5-10(k)?

- (e) Use the applet to estimate the p-value by using the **Count Samples** box: choosing to count “beyond” (for the two-sided p-value) and specifying the observed sample slope. Press **Count**.

You should see many similarities between these distributions (from random sampling and random shuffling). So let's again consider applying the *t*-probability model to the standardized statistics.

(f) Use the standard deviation for the slopes found in (d) and standardize the value of the observed sample slope.

(g) Now select the radio button for ***t*-statistic**. Use the applet to estimate the p-value corresponding to the observed *t*-statistic. How does it compare to the p-value you found in the previous Investigation?

(h) Check the box to **Overlay *t*-distribution**. Does it appear to be a reasonable model of the shuffled *t*-statistics distribution?

(i) Check the **Regression Table** box to confirm your calculations.

**Discussion:** It is interesting to compare these two analysis approaches. In Investigation 5.10, we modeled repeated random sampling from a population but this required us to make certain assumptions about the data in the population (which we will spell out in more detail in the next Investigation). In Investigation 5.11, we didn't make any assumptions about a larger population; instead we modeled what the sample data could look like if the observed response values had been randomly paired, over and over again, with the observed explanatory variable values. This approach is sometimes referred to as "conditioning on the observed data." It is not surprising that the results are similar for both approaches, but some consider the second approach to be more "flexible" and "less restrictive" in the assumptions it makes. However, the first approach allows for more elegant mathematical derivations predicting how the sampling distribution of sample slopes will behave.

If we had had a stronger association in our sample (and  $s$  and  $s_y$  were less similar in value), you would have seen more differences between these two approaches. In particular, the random shuffling will have a larger standard deviation of the sample slopes than the random sampling. We saw with random sampling that the variability in the slopes is controlled by  $\sigma$ , not by  $SD(Y)$ . However, with random shuffling, any  $x$ -value can be paired with any  $y$ -value, so the slopes are able to "swing around" a little more freely especially with smaller sample sizes. When you view the Regression Table, the  $SD(b_1)$  reported there will correspond to random sampling.

### Practice Problem 5.11

Confirm the preceding paragraph using the *height* and *footlength* data of Investigation 5.8.

**Investigation 5.12: Boys' Heights**

Suppose we have data on the height (in centimeters) of boys for 3 different random samples at ages 2, 3, 4. The data in [hypoHt.txt](#) are modeled after data from the Berkeley Guidance Study which monitored the height (and weight) of boys and girls born in Berkeley, California between January 1928 and June 1929.

- (a) Which is the explanatory variable and which is the response variable?

Explanatory variable:

Response variable:

- (b) Do you expect the 2-year-old boys to all have the same height? Do you expect the 3-year-old boys to all have the same height? Do you expect the mean height of the 3-year-old boys to be the same as the mean height of the 2-year-old boys? Explain.

The regression equation for these sample data turns out to be  $\hat{\text{height}} = 75.2 + 6.47\text{age}$ , with a correlation coefficient of  $r = 0.857$ .

- (c) Is it possible that in the *population* of all Berkeley boys in the 1920s, there really is no linear relationship between age (2–4 years) and height, but that we obtained a correlation coefficient and sample slope coefficient this large just by random chance? Explain.

- (d) Describe how we might investigate the likelihood of obtaining such sample statistics by chance alone, if there really were no association between age and height in the population.

One way to determine whether our sample slope is significantly different from a conjectured value of the population slope is to assume a mathematical model.

To understand the most basic regression model, consider conditioning your data set on a particular value of  $x$ . For example, let's look at the distribution of heights for each age group.

(e) Use the `hypotht` data to create stacked dotplots using the heights as the response variable and age as the grouping variable. Compare the shape, center, and variability of the three distributions:

How would you describe the shape of each distribution?

Numerically, how do the means differ for the three distributions? Has the average height increased roughly the same amount each year?

Is the variability in the heights roughly the same for each age group?

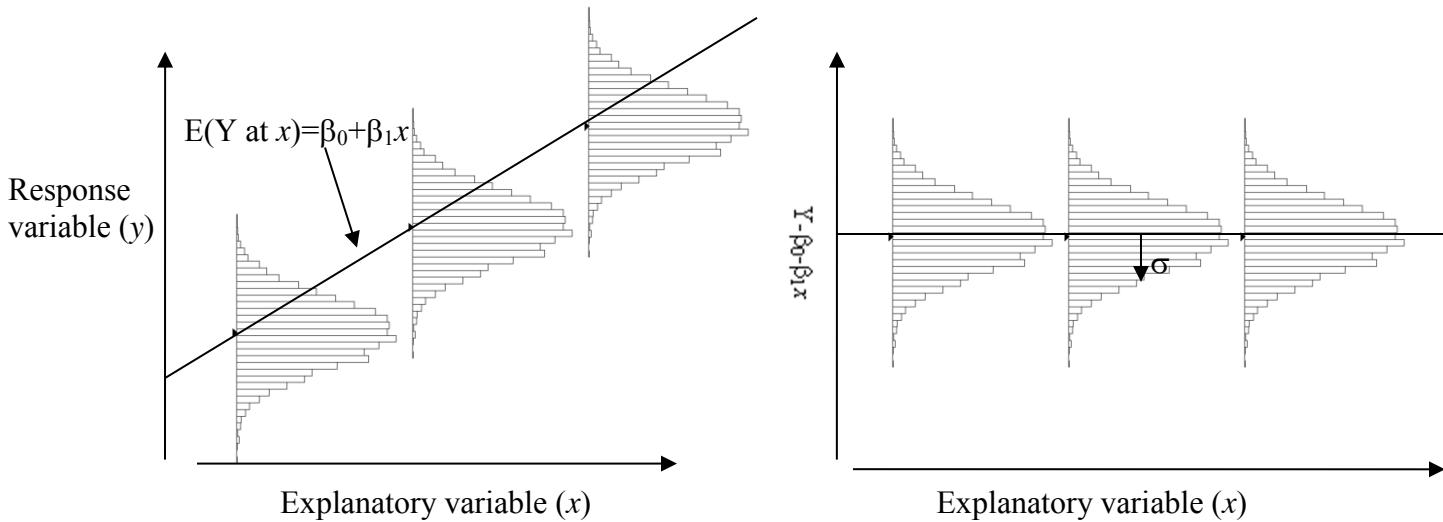
The [basic regression model](#) specifies the following conditions:

- The relationship between the mean (expected value) of the response variable and the explanatory variable is linear:  $E(Y \text{ at each } x) = \beta_0 + \beta_1 x$ .
- The variability of the observations at each value of  $x$ ,  $SD(Y \text{ at each } x)$ , is constant. We will refer to this constant value as  $\sigma$ . This means that the variability of the (conditional) response variable distributions does not depend on  $x$ .
- The distribution of the response variable at each value of  $x$  is normal.

To determine whether this mathematical model is appropriate we will check for three things: linearity, equal variance, and normality.

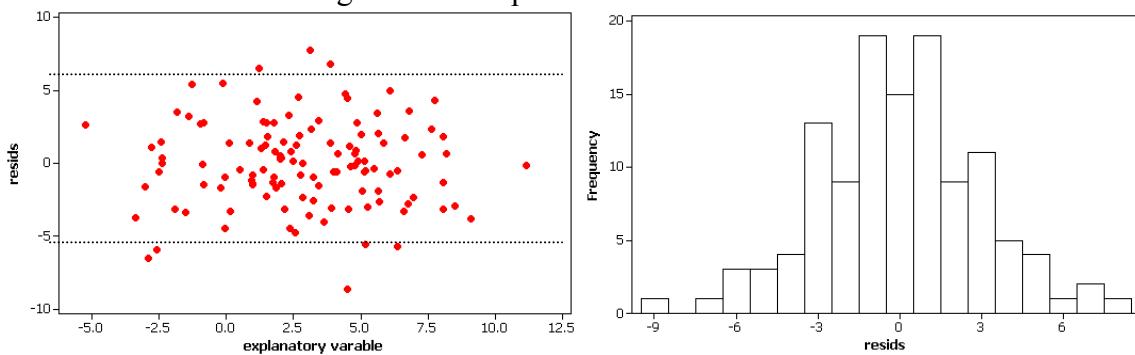
(f) Do these conditions appear to be met for the Berkeley boys' heights? Explain.

**Discussion:** In the above example, checking the conditions was straight forward because we could easily examine the distribution of the heights at the 3 different values of  $x$ . When we don't have many repeat values at each value of  $x$  we will need other ways to check these conditions. However, you should notice that the only thing that changes about the distribution of the response as we change  $x$  is the mean. So if we were to subtract the mean from each observation and pool these values together, then we should have one big distribution with mean 0, standard deviation equal to  $\sigma$ , and a normal shape.



However we don't know  $\beta_0$  and  $\beta_1$ , the population regression coefficients, but we have their least squares estimates. If we subtract the fitted values from each response value, these differences are simply the residuals from the least squares regression. So to check the technical conditions we will use *residual plots*.

- The linearity condition is considered met if a plot of the residuals vs. the explanatory variable does not show any patterns such as curvature.
- The equal variance condition is considered met if a plot of the residuals vs. the explanatory variable shows the similarly variability in residuals across the values of  $x$ .
- The normality condition is considered met if a histogram and normal probability plot of the residuals look normal.
- Another condition is that the observations are independent. We will usually appeal to the data being collected randomly or randomization being used. The main thing is to make sure there is not something like time dependence in the data.



The above graphs show data that satisfy these conditions nicely, and in this case we would consider the  $t$ -procedures from Investigation 5.10 to be valid.

**Investigation 5.13: Cat Jumping (cont.)**

Reconsider the data from Investigation 5.6 about factors that are related to a cat's jumping ability ([CatJumping.txt](#)).

(a) Use technology to determine the equation of the least squares line for predicting a cat's takeoff velocity from its percentage of body fat. Record the equation of this line, using good statistical notation.

(b) Produce a histogram and a normal probability plot of the residuals of this regression.

- In R You can access the residuals by using `> lm(price~sqft) $residuals`

Does the normality condition appear to be satisfied?

(c) Produce a graph of the residuals vs. the percentage of body fat variable. Does the equal variance condition appear to be met? Does the linearity condition appear to be met?

Equal variance:

Linearity:

(d) Consider testing whether the sample data provide strong evidence that percentage of body fat has a negative association with takeoff velocity in the population. State the hypotheses to be tested, and report (from the output) the value of the appropriate test statistic and p-value. Summarize your conclusion.

(e) Confirm that the *t*-test statistic for the hypotheses in (d) is equal to  $b_1/SE(b_1)$ .

(f) Use a *t*-procedure and the values of  $b_1$  and  $SE(b_1)$  to produce (by hand) a 95% confidence interval for the population slope  $\beta_1$ .

(g) Interpret the confidence interval that you produced in (f). Be sure to interpret not only the interval itself but also what the slope coefficient means in this context.

- (h) Use the equation of the least squares line to predict the takeoff velocity for a cat with 25 percent body fat. Then do the same for a cat with 50 percent body fat.

Predicted takeoff velocity with 25% body fat:

Predicted takeoff velocity with 50% body fat:

- (i) Which of the two predictions in (h) do you feel more comfortable with? Which do you suspect would be more *precise*? Explain.

Rather than only report one number as our prediction, we would like specify a confidence interval that indicates our “accurate” or precise we believe our prediction to be. The following procedure is valid if the basic regression model conditions are met.

### Technology Detour – Prediction Intervals

#### In R

Use the `predict` command, passing in the variables for the linear model but also a “new data” data frame for the value(s) you want predictions for (but giving them the same name as your explanatory variable). For example:

```
> predict(lm(velocity~percentbodyfat), newdata=
  data.frame(percentbodyfat=25), interval="prediction")
```

- (j) Report the 95% [prediction interval](#) for the takeoff velocity of a cat with 25% body fat from R/Minitab. Also determine the midpoint of this interval; does its value look familiar? Also interpret what this interval reveals.

Prediction interval:

Midpoint:

Interpretation:

- (k) Repeat (j) to obtain a 95% prediction interval for the takeoff velocity of a cat with 50% body fat. Which interval is wider? Is this what you predicted in (i)?

As you saw in Investigation 2.6, the level of precision will also depend on whether we want to predict the mean response or an individual response outcome.

**Definition:** Statistical packages compute both “prediction intervals” and “confidence intervals.” A prediction interval gives us the interval of plausible values for an individual response at a particular value of the explanatory variable. A confidence interval gives us the set of plausible values for the *mean* response at a particular value of the explanatory variable.

### Technology Detour – Confidence Intervals

**In R:** use `interval="confidence"`

**In Minitab:** Now report the 95% CI output.

(l) What does R/Minitab report for the 95% confidence interval for the average takeoff velocity among all cats that have 25% body fat? Also determine the midpoint of this interval; does its value look familiar? Also interpret what this interval reveals.

Confidence interval:

Midpoint:

Interpretation:

(m) How does the interval in (l) compare to the interval in (j)? Why does this relationship make sense? Explain.

**Study Conclusions**

Technology tells us that we are 95% confident that a cat with 25% body fat would have a takeoff velocity between 292.4 and 405.3 cm/sec. But if we were to consider the population of all cats with 25% body fat, we are 95% confident that the *mean* takeoff velocity of these cats is between 335.5 and 362.2 cm/sec, a much narrower interval. These procedures are valid because the analysis of the residual plots did not reveal any strong departures from the basic regression model conditions.

Note: As with other  $t$  procedures, these procedures are fairly robust to the normality condition if you have a larger sample size *except* the prediction interval calculation.

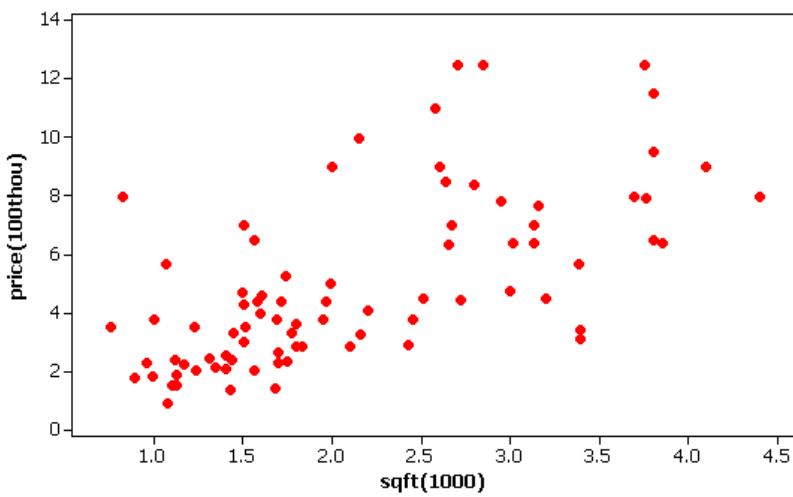
**Practice Problem 5.13**

Reconsider the 5K race results ([Talley5K2013.txt](#)) from Investigation 5.10.

- (a) Determine and interpret a 95% confidence interval for the mean finishing time of all 25-year-old runners in the population.
- (b) Joy Montoya, a 43-year-old female, did not appear in the first set of results. Use these data to predict her finishing time.
- (c) How do the intervals in (a) and (b) compare – Midpoint? Width? Explain why each has changed (or not) the way it has.
- (d) Repeat (b) and (c) for Laure James, a 53-year-old female who also didn't appear in the original results.
- (e) Do you consider the above interval calculations to be valid? (Examine and discuss residual plots.)

### Investigation 5.14: Housing Prices

A group of students wanted to investigate which factors influence the price of a house (Koester, Davis, and Ross, 2003). They used <http://www.househunt.com>, limiting their search to single family homes in California. They collected a stratified sample by stratifying on three different regions in CA (northern, southern, and central), and then randomly selecting a sample from within each strata. They decided to focus on homes that were less than 5000 square feet and sold for less than \$1.5 million. The file [housing.txt](#) contains data for their final sample of 83 houses. Below is a scatterplot of the size vs. selling price for this sample.



- (a) Open the `housing` file and determine the least-squares regression line for these data. Also report and interpret the value of  $r^2$  for these data.
- (b) Based on the scatterplot, do these data appear to fit the basic linear regression model?
- (c) Produce a histogram and a normal probability plot of the residuals of this regression. Is it reasonable to consider the residuals as following the normal distribution? Explain.
- (d) Produce a scatterplot of the residuals vs. the square footage.  
Does there appear to be curvature in this graph?  
Does the variability in the residuals appear to be roughly constant across the graph?

**Discussion:** It takes a while to become comfortable interpreting these residual plots, but the housing data do appear to have some problems. The residuals appear to cluster closer to zero for the smaller houses and to have more variability for the larger houses. This violates the condition of a constant standard deviation for each value of  $x$ . There is also a very slight hint of some curvature in this graph. The distribution of the residuals is clearly non-normal. This residual analysis gives us several reasons to *not* apply the basic regression model to these data.

When the regression model conditions are violated, we can often *transform* the data in the hopes that the new variables do meet the conditions.

(e) Take the log base 10 of both the house prices and sizes, storing the results as *logprice* and *logsqft*.  
[Hint: In Minitab, use Calc > Calculator or the `let` command. Also keep in mind that both R and Minitab assume natural log when use the “log” function. In Minitab use “logten” and in R use “log10”.] Produce the scatterplot and determine the regression equation to model the relationship between these two variables, storing the residuals. Examine the residual plots (histogram of residuals and residuals versus explanatory variable). Do these transformed variables appear to be more suitable for the basic regression model? Explain.

(f) Interpret the slope coefficient in the regression model using the log transformed variables.

(g) Use the least squares regression line to predict the cost of a 3,000 square foot house. [Hints: Substitute  $\log_{10}(3000)$  into the right hand side of the equation to obtain a prediction for the  $\log_{10}(\text{price})$ , and then raise 10 to this power to determine the predicted price. Similarly, you can back transform the endpoints of the confidence and prediction intervals from R/Minitab.]

**Discussion:** The log transformation has made improvements for both the normality condition and the equal variance condition. It is not uncommon for one transformation to “correct” more than one problem. We could continue to explore other transformations (e.g., square root and power transformations) until we find transformed data that are more appropriate for the basic regression model. Although these transformations are useful, we then have to be careful in how we interpret the regression coefficients and make predictions. If possible, it is usually most useful to “back-transform” to the original scale.

### Practice Problem 5.14

The datafile [walmart.txt](#) contains data on the number of stores and the number of SuperCenters that Wal-Mart had in operation between 1989 and 2002.

- Create a scatterplot of the number of SuperCenters that Wal-Mart has had in operation versus time.
- Would it be appropriate to use the basic regression model for these data? Explain.
- Transform the SuperCenters variable by taking the square root. Would it be appropriate to use the basic regression model for the relationship between this variable and year? Explain.
- Transform the SuperCenters variable by taking the natural log. Would it be appropriate to use the basic regression model for the relationship between this variable and year? Explain.
- Choose the best model from either (c) or (d) (and justify your choice), find the least-squares regression equation and use it to predict how many SuperCenters Wal-Mart had in operation in 2003.
- Report and interpret a 95% prediction interval for 2003.

### Summary of Inference for Regression

To test  $H_0: \beta_1 = \text{hypothesized slope}$  vs.  $H_a: \beta_1 \neq \text{hypothesized slope}$

The test statistic:  $t = \frac{b_1 - \text{hypothesized slope}}{SE(b_1)}$  follows a  $t$ -distribution with  $df = n - 2$

An approximate  $100 \times C\%$  confidence interval formula is  $b_1 \pm t_{n-2} SE(b_1)$  where  $-t_{n-2}$  is the  $100 \times (1 - C)/2$  percentile of the student  $t$ -distribution with  $n - 2$  degrees of freedom.

These procedures are valid as long as

- L: There is a linear relationship between the response and explanatory variable.
- I: The observations are independent.
- N: The response follows a normal distribution for each value of  $x$ .
- E: The standard deviation of the responses is the equal at each value of  $x$ .

These conditions are checked by examining the residual plots: residuals vs. explanatory variable and histogram/normal probability plot of the residuals.

The Analyzing Two Quantitative Variables applet, Minitab ([Stat > Regression > Regression > Fit Regression Model](#)), and R (`>summary(lm(response~explanatory))`) all automatically report the *two-sided* p-value. This p-value can be divided in half to obtain a one-sided p-value (assuming the observed slope is in the direction conjectured).

**Discussion:** Typically we are most interested in testing whether the population slope coefficient equals zero, as that would suggest that the population regression line is flat and so the explanatory variable is of no use in predicting the response variable. Because testing whether the slope coefficient is zero tests whether the regression model with this explanatory variable is of any use, it is sometimes called a *model utility test*.

There are situations where we might want to test other values of the population slope, for example  $\beta_1 = 1$  (do  $x$  and  $y$  change at the same rate?). We can also use the same approach to carry out tests and create confidence intervals for the population intercept. However, these are less useful in practice, in particular because the intercept is often far outside the range of data and does not have a meaningful interpretation.

Checking the technical conditions is an important step of the process. Notice that you will typically fit the regression model first to obtain the residuals to allow you to assess the model. As with ANOVA, these  $t$  procedures are fairly robust to departures from normality. If the conditions are violated, you can again explore transformations of the data.

## Technology Exploration: The Regression Effect

In the 2013 Masters golf tournament, Spaniard Sergio Garcia had a score of 66 in the first round but a score of 76 in the second round. Keep in mind that in golf, lower scores are better than higher scores. Is this evidence that he choked? The data file [Masters13.txt](#) contains golfers' scores in the first two rounds of the 2013 Masters.

(a) Use technology to construct a scatterplot of second round score vs. first round score. Also calculate the correlation between these two. Does it reveal an association between scores in these two rounds? If so, is it positive or negative? Would you call it weak, moderate, or strong?

(b) Use technology to sort the data according to the golfers' first-round scores:

- In R
  - > mydata=data.frame(golfer, round1, round2)
  - > sortdata=mydata[order(mydata[,2]),]

List the ten golfers who scored lowest in the first round, along with their scores in both rounds. Then list the ten golfers who scored highest in the first round, along with their scores in both rounds.

(c) How many of the “top ten” from the first round improved (scored lower) in the second round than the first? How many of the “bottom ten” from the first round improved?

(d) Which group saw more people improve? Is this difference small or substantial?

(e) Determine the median second round score for the two groups (the “top ten” from round one and the “bottom ten” from round one). Which group tended to score better (lower) in the second round?

(f) Add the “ $y = x$ ” line to your scatterplot.

- In R: > lines(round1, round1)

Identify any golfers by name who improved in the second round after scoring in the top ten of the first round. Identify any golfers by name who did worse in the second round after scoring in the bottom ten of the first round.

(g) Do you think the “ $y = x$ ” line does a good job of summarizing the relationship?

How do you think the regression line will compare? Will the slope be greater or smaller?

(h) Add the  $\bar{y}$  line to the scatterplot.

- In R: > abline(h=mean(round1))

Do you think this line does a good job of summarizing the relationship? How do you think the regression line will compare? Will the slope be greater or smaller?

(i) Now add the regression line to the scatterplot.

- In R: > abline(lm(round1~round2), col=3)

Where does the regression line fall compared to the two earlier lines?

(j) Determine the regression equation. Is the slope less than one? Explain why this is to be expected whenever the two standard deviations are similar.

[Hint: Recall that the slope coefficient can be expressed as  $b = r(s_y)/(s_x)$ .]

(k) Explain why your answer to (j) suggests that golfers with poor first round scores will tend to improve but golfers with good first round scores will tend to worsen.

*This phenomenon is sometimes called the regression effect or regression to the mean. Sir Francis Galton first discussed this effect in the late 1880s when he examined data on the heights of parents and children, calling it regression to mediocrity.*

## Section 5.4 Summary

You continued to study least squares regression in this section, turning to issues of *inference*. You learned the conditions of the basic regression model (linearity, independence, normality, and equal variance), and once again you used simulation to study the approximate sampling distribution assuming the null hypothesis is true, this time for sample regression coefficients. You found that the *t*-distribution can be used to perform significance tests and construct confidence intervals for regression coefficient parameters and for predicted values, based on sample data. The primary significance test of interest is usually whether the sample data provide evidence of an association in the population between the two variables, which is equivalent to testing whether the population slope coefficient equals zero. You also learned how to use residual plots to check the technical conditions associated with the basic regression model. Example 5.3 illustrates an application of regression analysis.

### Example 5.1: Internet Use by Region

*Try these questions yourself before you use the solutions following to check your answers.*

The Pew Internet and American Life Project examines how Americans use the internet. In 2002 the organization took random samples of people from across the country and asked questions about their use of the internet. Consider the following information from this study:

|                     | Northeast | South | Midwest | West |
|---------------------|-----------|-------|---------|------|
| Sample size         | 3973      | 4332  | 4929    | 5137 |
| # of internet users | 2417      | 2372  | 2831    | 3259 |

Analyze these data to address the question of whether internet use varies across these four regions of the country. Include graphical and numerical summaries as well as a test of significance. Summarize your conclusions.

## Analysis

We will treat the samples taken by the Pew organization as independent random samples from these four regions of the country. With a binary response variable (internet user or not) and more than two groups to compare, we will use a chi-square analysis.

Let  $\pi_i$  represent the actual population proportion of internet users in region  $i$ . Then the hypotheses to be tested are:

$$H_0: \pi_{NE} = \pi_S = \pi_{MW} = \pi_W$$

(The population proportions of internet users are the same for all four regions.)

$H_a$ : The population proportion of internet users is different in at least one region.

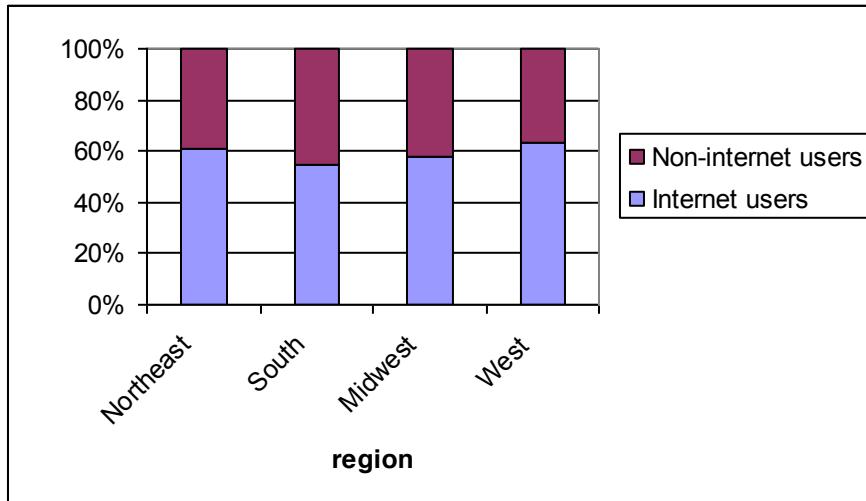
The two-way table of observed counts is:

|                    | Northeast | South | Midwest | West | Total  |
|--------------------|-----------|-------|---------|------|--------|
| Internet users     | 2417      | 2372  | 2831    | 3259 | 10,879 |
| Non-internet users | 1556      | 1960  | 2098    | 1878 | 7492   |
| Total              | 3973      | 4332  | 4929    | 5137 | 18,371 |

The sample proportions of internet users in the four regions are:

|                              | Northeast | South | Midwest | West | Total |
|------------------------------|-----------|-------|---------|------|-------|
| Proportion of internet users | .608      | .548  | .574    | .634 | .592  |

A segmented bar graph to display these data is:



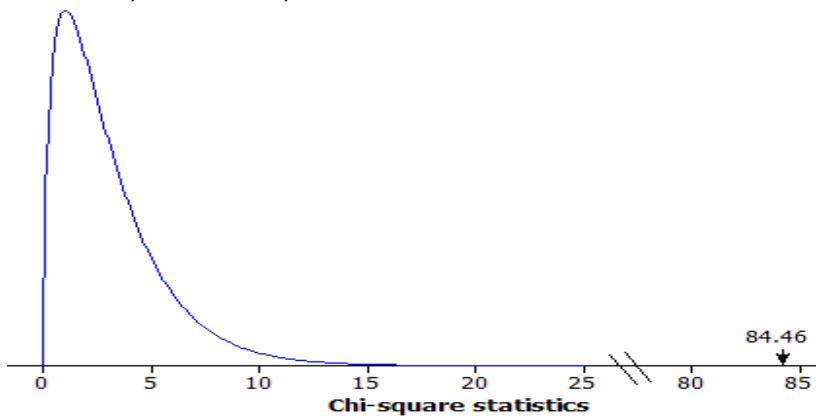
We notice that the proportions of internet users do not vary too much across these four regions. The west has the highest proportion, with about 63% internet use, and the south has the smallest with only 55% internet use.

To see whether at least one of these regions is statistically significant different from the others, we use technology to determine the expected counts, chi-square test statistic, and p-value. Below is Minitab output:

Expected counts are printed below observed counts  
 Chi-Square contributions are printed below expected counts

|       | Northeast | South   | Midwest | West    | Total |
|-------|-----------|---------|---------|---------|-------|
| 1     | 2417      | 2372    | 2831    | 3259    | 10879 |
|       | 2352.74   | 2565.34 | 2918.87 | 3042.05 |       |
|       | 1.755     | 14.571  | 2.645   | 15.473  |       |
| 2     | 1556      | 1960    | 2098    | 1878    | 7492  |
|       | 1620.26   | 1766.66 | 2010.13 | 2094.95 |       |
|       | 2.548     | 21.158  | 3.841   | 22.468  |       |
| Total | 3973      | 4332    | 4929    | 5137    | 18371 |

Chi-Sq = 84.460, DF = 3, P-Value = 0.000



This procedure is valid because all of the expected counts are far larger than 5, and we are assuming that the Pew organization took independent random samples across the regions. Their report mentions that they used random digit dialing with extensive follow-up efforts, but it's not clear whether they took independent samples within these regions or took a nationwide sample and treated region as one of the variables. Either way the chi-square analysis is valid.

The p-value is extremely small (0.000 to three decimal places), indicating that the differences observed in the sample proportions would almost never occur by chance if the population proportions were really equal across the four regions. Thus, we have overwhelming evidence to reject the null hypothesis and conclude that the proportions of internet users are not the same in these four regions.

In an effort to say a little bit more, we can examine the cell contributions. We see that the largest contributions occur in the West and South. The West had more internet users, 3259, (and so fewer non-internet users) than would have been expected if the proportions were all the same, 3042.05, and the South had fewer internet users, 2372, (and so more non-users) than would have been expected, 1766.66. The observed counts in the Northeast and Midwest regions were fairly close to the expected counts.

In conclusion, the sample data provide very strong evidence that the population proportions of internet users differ across regions of the United States. The differences in the sample proportions are larger than can reasonably be explained by random sampling variation. The large sample sizes in this study help to render the differences among the samples statistically significant even though the sample proportions seem to be fairly similar in practical terms.

**Example 5.2: Lifetimes of Notables**

The 1991 *World Almanac and Book of Facts* contained a section on “noted personalities” in a total of nine occupational categories. The lifetimes of these people can be found in the [lifetimesFull.txt](#) data file. Analyze these data to address the question of whether the average lifetimes across these occupational categories differ significantly. Include graphical and numerical summaries as well as a test of significance. Summarize your conclusions.

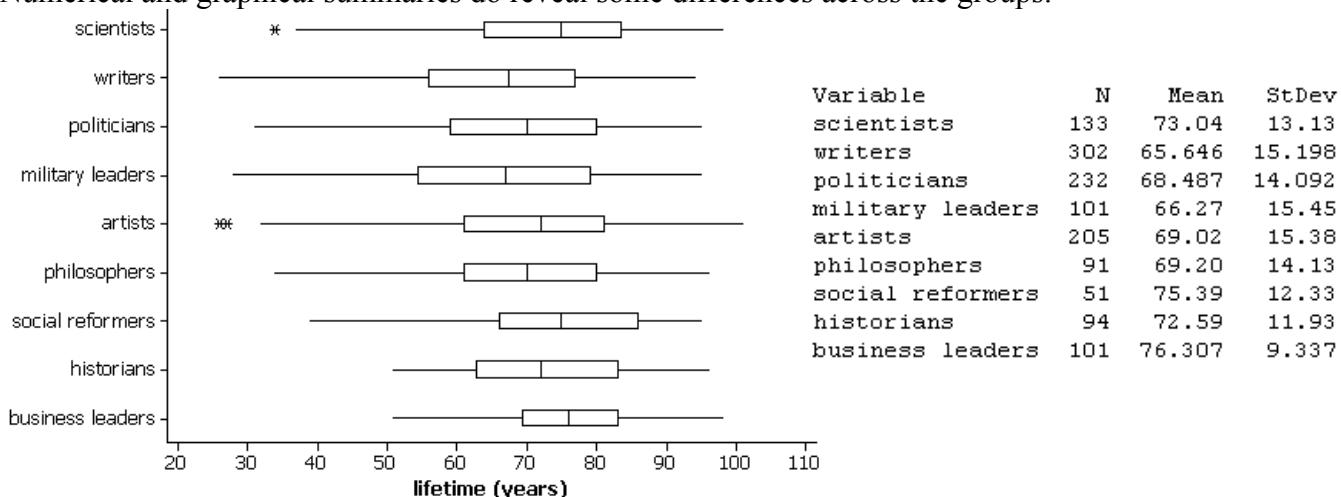
## Analysis

To compare several groups on a quantitative variable, we will consider analysis of variance. The first step will be to describe the data we have. The file `lifetimesFull` contains the lifetimes (years) of notable personalities from nine different occupation classifications. The samples were selected independently from within each job classification but they were presumably not selected at random. Thus, we should be cautious in how we generalize our conclusions. We will need to restrict our conclusions to “notable personalities” but otherwise might not suspect any bias in the *World Almanac’s* sampling method. In this case, we can define  $\mu_i$  to be the population mean lifetime for all notable personalities with occupation  $i$ .

$$H_0: \mu_{\text{scientist}} = \mu_{\text{writers}} = \mu_{\text{politicians}} = \mu_{\text{military}} = \mu_{\text{artists}} = \mu_{\text{philosophers}} = \mu_{\text{social}} = \mu_{\text{historians}} = \mu_{\text{business}}$$

$$H_a: \text{at least one population mean lifetime differs from the rest}$$

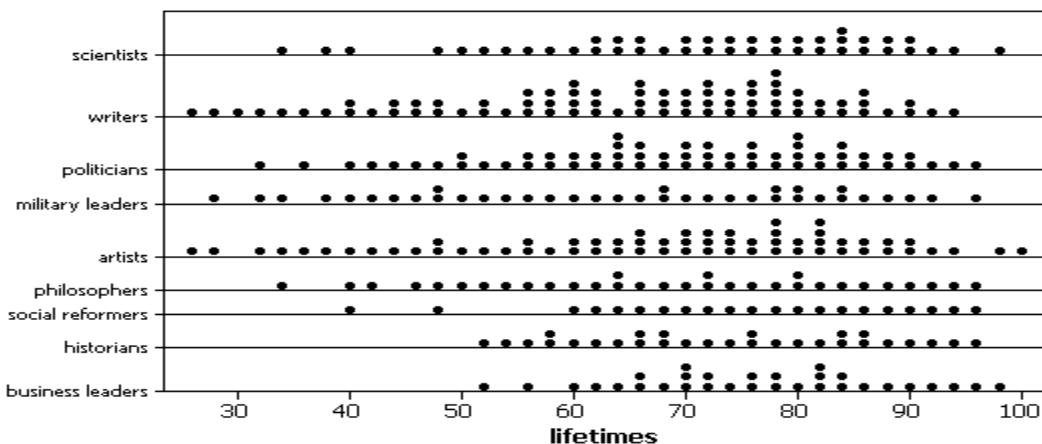
Numerical and graphical summaries do reveal some differences across the groups.



There is a slight tendency for longer lifetimes among scientists and business leaders and for shorter lifetimes among writers and military leaders. However, there is a fair bit of variation within the groups and much overlap in the boxplots.

To decide whether an ANOVA procedure is valid, we check the technical conditions:

- Dotplots of the individual samples do not reveal any marked departures from normality.

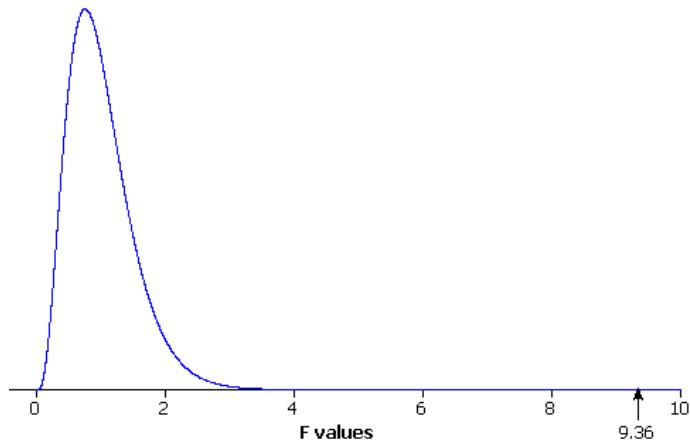


- The ratio of the largest to smallest standard deviations ( $15.45/9.34$ ) is less than 2 so we will assume that the population variances are equal.
- We have independent samples from each population though the randomness condition is questionable.

Although we do not have randomness in this study, we will still explore whether the difference between the sample means is larger than we would expect by chance.

#### One-way ANOVA: lifetime versus category

| Source   | DF   | SS     | MS   | F    | P     |
|----------|------|--------|------|------|-------|
| category | 8    | 14837  | 1855 | 9.36 | 0.000 |
| Error    | 1301 | 257818 | 198  |      |       |
| Total    | 1309 | 272655 |      |      |       |



The large test statistic  $F = 9.36$  and small p-value ( $0.000 < 0.001$ ) provide strong evidence that the population mean lifetime for at least one of these occupations differs from the rest.

Because this was an observational study, we cannot draw any cause and effect conclusions and as discussed above we should be cautious in generalizing these conclusions beyond the 1310 individuals in the study as they were not randomly selected. For the individuals in this study, there appears to be something other than random chance to account for the differences we observed in their lifetimes.

**Example 5.3: Physical Education Class Performance**

A physical education teacher at a junior high school in Central California wanted to determine whether there is a relationship between seventh-graders' times to run a mile and how many push-ups they can do under controlled conditions. She collected the data in [PEClass.txt](#) as part of the mandated state physical fitness testing program. Analyze these data to address the question of whether they provide evidence of a relationship between these two variables. Include graphical and numerical summaries as well as a test of significance. Summarize your conclusions.

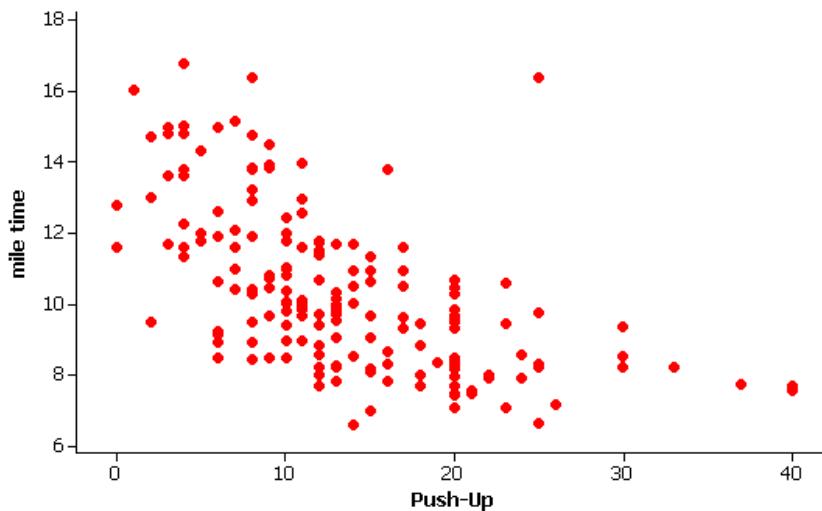
## Analysis

Because we have two quantitative variables here (time to run one mile and number of push-ups), the appropriate analysis would be regression. We first want to examine numerical and graphical summaries. When you open PEClass you will notice that the mile run times have been recorded in “time format.” We want to convert this to numerical values. For example, in Minitab choose Data > Change Data Type > Data/Time to Numeric. Specify C3 as the column to be changed and C4 as the storage location for the converted values. The data are now numeric, but in terms of a 24-hour day. To convert these back to minutes, type

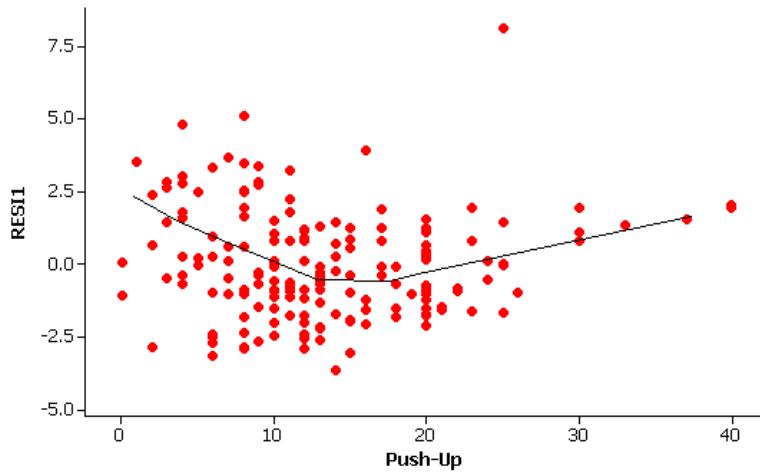
MTB> let c5=c4\*24 or use Calc > Calculator, letting *miletime* = 24\*c4.

Now you should have the number of minutes (including the fraction of minute) for each student.

Because we aren’t considering either of these as a response variable, it does not matter which variable we denote as the *y*-variable and which as the *x*-variable. If we plot mile time vs. push-ups, we see there is a strong negative association.

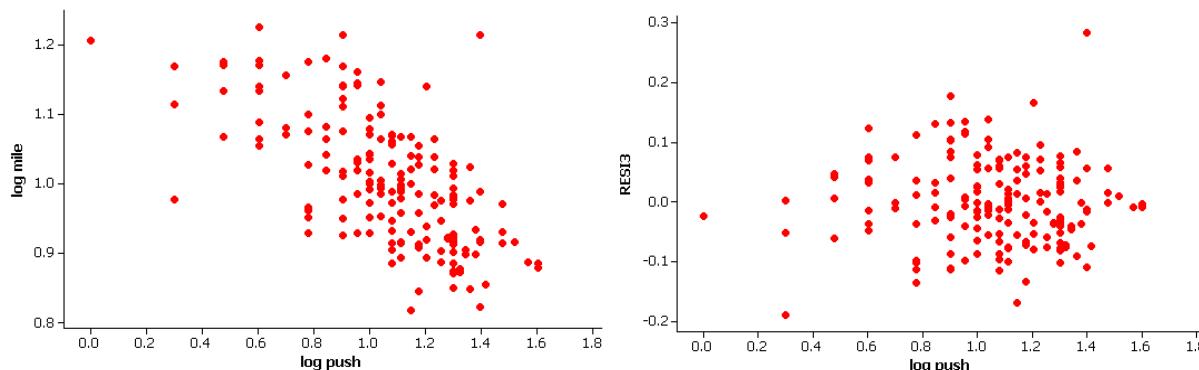


Students who do more push-ups also tend to run the mile in faster times. However, there is some evidence that the relationship is not linear. There is also an unusual observation, a student who did a larger number of push-ups but was one of the slowest runners. Carrying out the regression and examining residual plots confirms these observations.

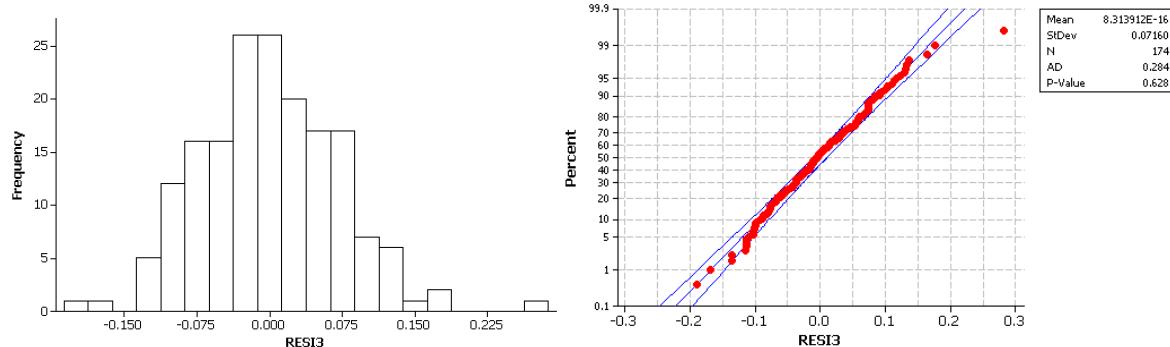


The residual vs. explanatory variable graph also reveals some differences in the amount of variation in the residuals at different values of the explanatory variable (indicating a violation of the constant variance condition).

It appears that transforming these data might be helpful. Because both distributions appear skewed to the right (if you looked at histograms of each variable individually), we could consider taking the log of each variable. The log-log scatterplot does appear more well behaved. (We have used log base ten but natural logs would also work.)



We still have some outliers but appear to now have a linear relationship. If we also examine the normality of the residuals:



This condition also seems to be reasonably met for the transformed data. There is slight evidence of skewness to the right in the residuals but coupled with the large sample size we will not be concerned with this minor deviation.

The correlation coefficient for the transformed variables is  $-0.624$ , indicating a moderately strong negative linear relationship between  $\log(\text{time})$  and  $\log(\text{push-ups})$ . The least-squares regression equation is computed by Minitab to be  $\hat{\log \text{mile}} = 1.23 - 0.213 \log(\text{push-ups})$ . The intercept coefficient here would indicate the predicted  $\log(\text{time})$  for a student who only completes 1 push up (so  $\log(\text{push-ups}) = 0$ ) to be 1.23. This corresponds to a time of  $10^{1.23}$  or about 17 minutes. The slope coefficient predicts the average multiplicative change in the  $\log(\text{time})$  times for each unit increase in  $\log(\text{push-ups})$ . A unit increase in " $\log(\text{push-ups})$ " corresponds to the push-ups increasing by a factor of 10. So for each 10 fold increase in the number of push-ups (e.g., 1 push up to 10 push-ups), the mile time decreases on average by a factor of  $10^{-0.213} = 0.61$ . [Note: our prediction for 10 push-ups is 1.017, corresponding to  $10^{1.017} \approx 10.4$  minutes, which is 0.61(17).]

If we test the significance of this correlation:

Let  $\beta$  represent the true population slope between  $\log(\text{time})$  and  $\log(\text{push-ups})$

$H_0: \beta = 0$  (there is not association between these two variables)

$H_a: \beta \neq 0$  (there is an association)

We find a test statistic of  $t = -10.48$  and a two-sided p-value of approximately 0.

| Predictor | Coef     | SE Coef | T      | P     |
|-----------|----------|---------|--------|-------|
| Constant  | 1.24825  | 0.02218 | 56.28  | 0.000 |
| log push  | -0.21315 | 0.02033 | -10.48 | 0.000 |

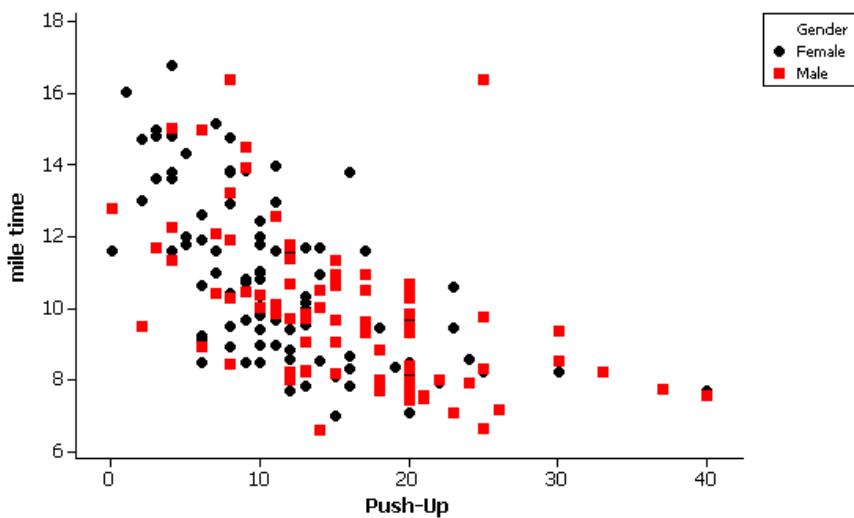
Note: This p-value is the same as reported by Minitab with the sample correlation coefficient.

With such a small p-value ( $< 0.001$ ) we would consider the relationship between  $\log(\text{time})$  and  $\log(\text{push-ups})$  to be statistically significant. Although we need to have some caution in generalizing these results to other schools, we can eliminate “random chance” as an explanation for the strong log-log relationship observed in this sample.

If we wanted to use this model to carry out predictions, we need to keep the transformed nature of the variables in mind. For example, if a student completed 25 push-ups during that portion of the test, we would predict  $1.23 - 0.213 \log_{10}(25) = 0.932$  for the log-time and therefore  $10^{0.932} = 8.56$  minutes for the mile time. We also need to start being a little cautious in predicting the mile time for such a large number of push-ups as we do not have a large amount of data in this region and our estimate will not be as precise.

In summary, there is a strong negative relationship between number of push-ups and time to run a mile for this 7<sup>th</sup> and 8<sup>th</sup> graders as we would expect (those students who do more than the average number of push-ups will tend to be the same students who complete a mile faster, in a below average time). This relationship can be considered statistically significant after performing log transformations, however, we have to be cautious in generalizing these results beyond this sample as the students were not randomly selected from a larger population of junior high students. Because this is an observational study, we are not claiming that doing more push-ups will cause the mile time to decrease.

We also saw in an earlier example that there was a statistically significant difference between males and females on these tasks and we might want to consider incorporating that variable into our analysis as well. For example, a coded scatterplot shows that the males tended to do more push-ups and also tended to be slightly faster than females on the mile times.



We can actually carry out a *multiple regression analysis* to predict *mile time* from both *push ups* and *gender*.

#### Coefficients

| Term     | Coef      | SE Coef   | T       | P     |
|----------|-----------|-----------|---------|-------|
| Constant | 0.529616  | 0.0125671 | 42.1430 | 0.000 |
| Push-Up  | -0.007340 | 0.0008250 | -8.8977 | 0.000 |
| Gender   |           |           |         |       |
| Female   | 0.000481  | 0.0061047 | 0.0788  | 0.937 |

This analysis reveals that push-ups are strongly related to mile times even after adjusting for gender (meaning this association is strong among the females and also among the males), but if you know the someone's push-up count, it is not that valuable to also know their gender.

**Example 5.4: Comparing Popular Diets**

Dansinger, Griffith, Gleason et al. (2005) report on a randomized, comparative experiment in which 160 subjects were randomly assigned to one of four popular diet plans: Atkins, Ornish, Weight Watchers, and Zone (40 subjects per diet). These subjects were recruited through newspaper and television advertisements in the greater Boston area; all were overweight or obese with body mass index values between 27 and 42. Among the variables measured were

- Which diet the subject was assigned to
- Whether or not the subject completed the twelve-month study
- The subject's weight loss after two months, six months, and twelve months (in kilograms, with a negative value indicating weight gain)
- The degree to which the subject adhered to the assigned diet, taken as the average of 12 monthly ratings, each on a 1-10 scale (with 1 indicating complete non-adherence and 10 indicating full adherence)

Data for the 93 subjects who completed the 12-month study are in the file [ComparingDiets.txt](#).

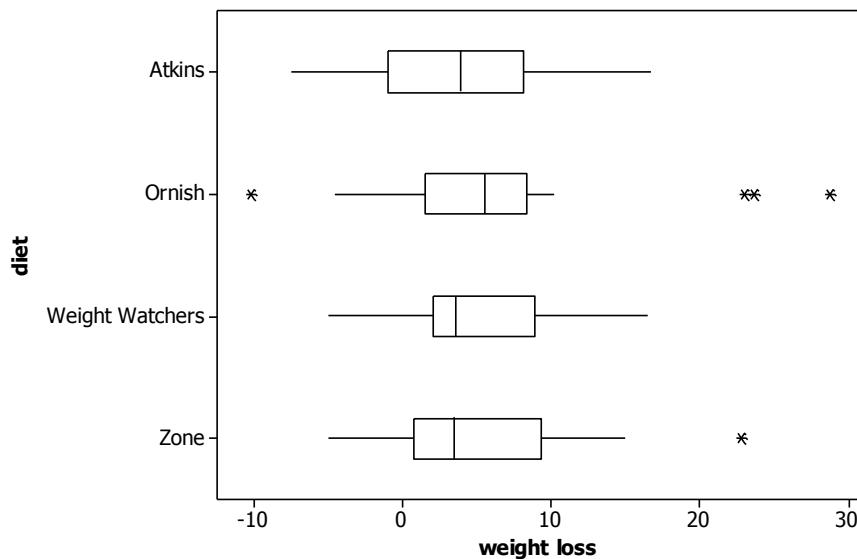
Some of the questions that the researchers studied are:

- (a) Do the average weight losses after 12 months differ significantly across the four diet plans?
- (b) Is there a significant difference in the completion/dropout rates across the four diet plans?
- (c) Is there a significant positive association between a subject's adherence level and his/her amount of weight loss?
- (d) Is there strong evidence that dieters actually tend to lose weight on one of these popular diet plans?

For each of these research questions, first identify the explanatory variable and the response variable, and classify each as categorical or quantitative. Then use graphical and numerical summaries to investigate the question, and summarize your findings. Next, identify the inference technique that can be used to address the question, and apply that technique. Be sure to include all aspects of the procedure, including a check of its technical conditions. Finally, summarize your conclusions for each question. Write a paragraph summarizing your findings from these four analyses. [Hint: To determine the completion rate for each diet, count how many of the 93 subjects who completed the study are in each diet group and compare those counts to the 40 that were originally assigned to each diet.]

## Analysis

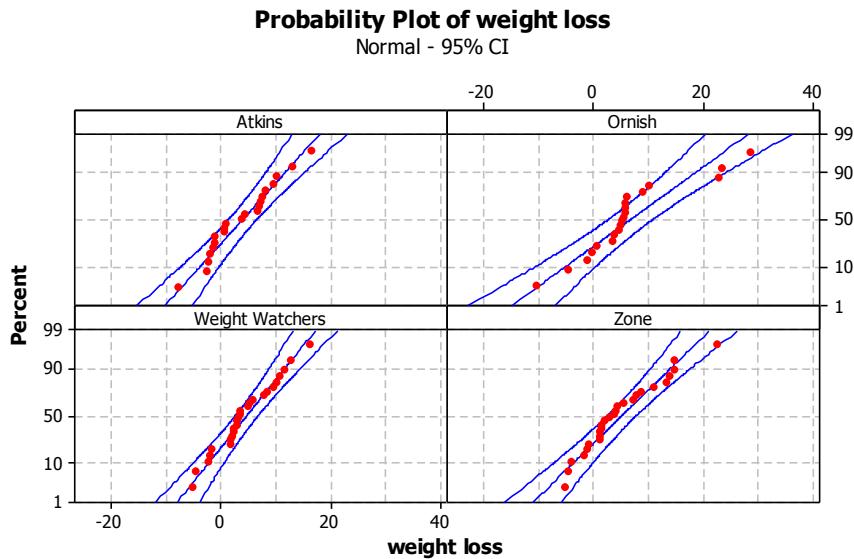
(a) Effect of diet plan on average weight loss: The explanatory variable is diet plan, which is categorical. The response variable is amount of weight loss after 12 months, which is quantitative. Boxplots and numerical summaries follow:



| diet      | N  | Mean | StDev | Median | IQR  |                    |
|-----------|----|------|-------|--------|------|--------------------|
| Atkins    | 21 | 3.92 | 6.05  | 3.90   | 9.15 | (all in kilograms) |
| Ornish    | 20 | 6.56 | 9.29  | 5.45   | 6.81 | (all in kilograms) |
| Wgt Watch | 26 | 4.59 | 5.39  | 3.60   | 6.85 | (all in kilograms) |
| Zone      | 26 | 4.88 | 6.92  | 3.40   | 8.63 | (all in kilograms) |

These boxplots and statistics seem to indicate that the four diets do not differ substantially with regard to weight loss after 12 months. The mean and median weight loss are both positive for all four diets, indicating that subjects did tend to lose some weight on these diets, roughly 4-6 kilograms on average. The boxplots also show substantial overlap between the four distributions. The means and medians are very similar for three of the diets, with the Ornish diet having a somewhat larger mean and median weight loss (6.56 and 5.45 kilograms, respectively) than the others. All four distributions of weight loss appear to be fairly symmetric, perhaps a bit skewed to the right. The variability in weight losses is also similar across all four diet plans, with the Ornish diet having the most variability, largely due to its one small and three large outliers.

Because we have a categorical explanatory variable and a quantitative response variable, we will apply ANOVA to these data. The technical conditions appear to be met: The subjects were randomly assigned to diet plans, the distributions look fairly normal (see the following normal probability plots), and the standard deviations are similar (ratio of largest to smallest is  $9.29/5.39$ , which is less than 2).



The hypotheses are:

- $H_0: \mu_A = \mu_O = \mu_W = \mu_Z$ , where  $\mu_i$  represents the underlying treatment mean weight loss after 12 months with diet  $i$ . This hypothesis says that the treatment mean is the same for all four diets.
- $H_a$ : that at least two of the treatment means differ; in other words, that at least one diet does have a different treatment mean than the others

Minitab produces the following ANOVA table:

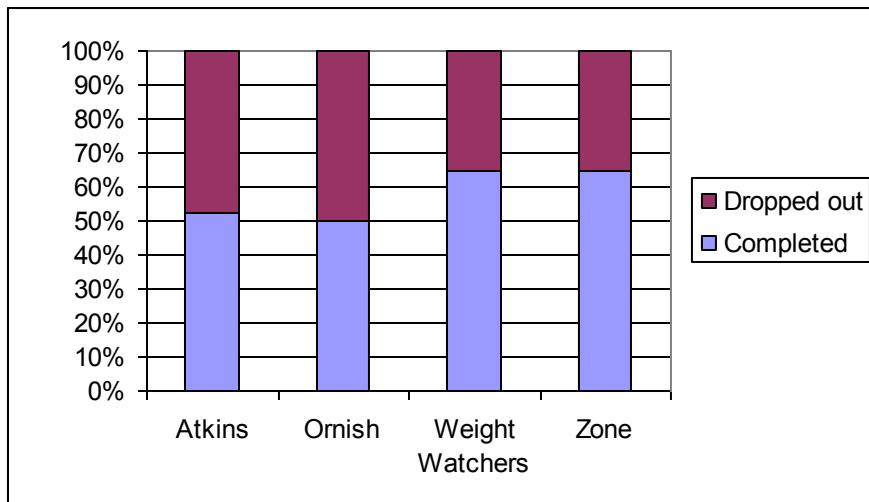
| Source | DF | SS     | MS   | F    | P     |
|--------|----|--------|------|------|-------|
| diet   | 3  | 77.6   | 25.9 | 0.54 | 0.659 |
| Error  | 89 | 4293.7 | 48.2 |      |       |
| Total  | 92 | 4371.3 |      |      |       |

The small  $F$ -statistic ( $F = 0.54$ ) and large p-value (0.659) reveals that the experimental data provide essentially no evidence against the null hypothesis. The p-value reveals that differences among the group means at least as big as those found in this experiment would occur about 66% of the time by randomization alone even if there were no true differences among the diets. In other words, the treatment means do not differ significantly, and there is no evidence that these four diets produce different average amounts of weight loss.

(b) Effect of diet plan on completion rate: The explanatory variable is diet plan, which is categorical. The response variable is whether the subject completed the study or dropped out, which is also categorical.

The two-way table of completion/dropout status by diet plan, followed by completion proportions and a segmented bar graph:

|                              | Atkins | Ornish | Weight Watchers | Zone |
|------------------------------|--------|--------|-----------------|------|
| <b>Completed</b>             | 21     | 20     | 26              | 26   |
| <b>Dropped out</b>           | 19     | 20     | 14              | 14   |
| <b>Completion proportion</b> | .525   | .500   | .650            | .650 |



This preliminary analysis appears to reveal that the completion rates are very similar across the four diet plans. Weight Watchers and Zone tied for the highest completion rate (62.5%), with Ornish having the lowest completion rate (50%), but these do not seem to differ substantially.

To test whether these differences in the distributions of the categorical response variable are statistically significant, we can apply a chi-square test of the hypotheses:

$H_0: \pi_A = \pi_O = \pi_W = \pi_Z$ , where  $\pi_i$  represents the underlying completion rate after 12 months with diet  $i$   
(diet does not have an effect on completion rate)

$H_a$ : at least two of the underlying completion rates differ (there is a difference in underlying completion rates across the 4 diets)

Minitab produces the following output:

|       | Weight |        |          |       |       |
|-------|--------|--------|----------|-------|-------|
|       | Atkins | Ornish | Watchers | Zone  | Total |
| 1     | 21     | 20     | 26       | 26    | 93    |
|       | 23.25  | 23.25  | 23.25    | 23.25 |       |
|       | 0.218  | 0.454  | 0.325    | 0.325 |       |
| 2     | 19     | 20     | 14       | 14    | 67    |
|       | 16.75  | 16.75  | 16.75    | 16.75 |       |
|       | 0.302  | 0.631  | 0.451    | 0.451 |       |
| Total | 40     | 40     | 40       | 40    | 160   |

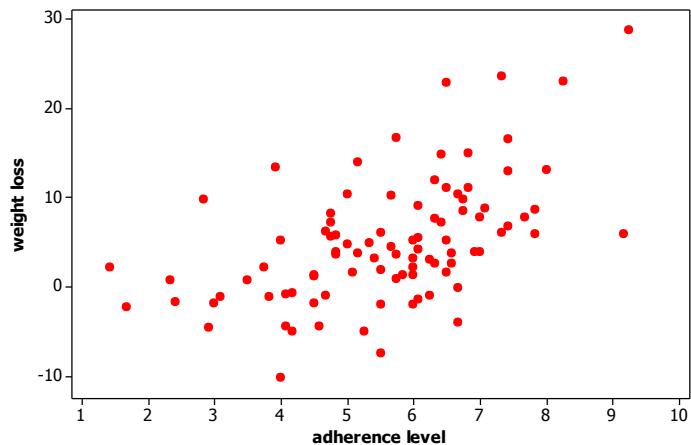
Chi-Sq = 3.158, DF = 3, P-Value = 0.368

Checking the technical conditions of the chi-square procedure, we note that the subjects were randomly assigned to a diet plan group and that all expected counts in the table are larger than 5 (smallest is 16.75), so we are justified in applying the chi-square test. The p-value of 0.368 says that if there were no difference in the underlying completion rates (i.e., no treatment effect) of completion among the four diet plans, then it would not be surprising (probability 0.368) to obtain experimental completion proportions that differ as much as these do. Because this p-value is not small, we can conclude only that

the experimental data do not provide evidence to suggest that the completion proportions differ across these four diet plans.

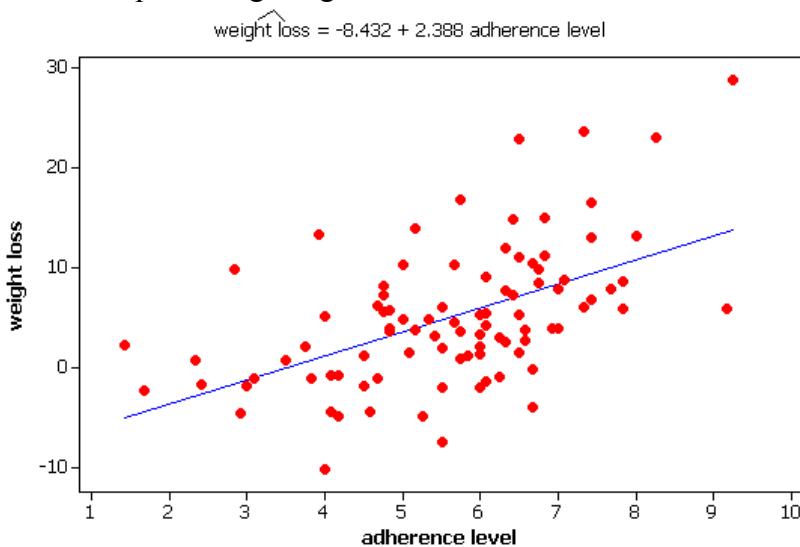
(c) Association between adherence and weight loss: The explanatory variable is adherence level. This variable is quantitative. (It could be considered categorical if it were simply on a 1–10 scale, but it is the average of 12 such values and so should be treated as quantitative.) The response variable is weight loss, which is quantitative.

A scatterplot of weight loss vs. adherence level follows:



This graph reveals a moderately strong, positive, linear relationship between weight loss and adherence level. The correlation coefficient can be found to be  $r = 0.518$ . The scatterplot and correlation coefficient both suggest that there is a positive association between these variables, that subjects with higher adherence levels tend to lose more weight.

We can fit a regression line for predicting weight loss from adherence level:



This model indicates that for each additional step on the adherence level scale, the subject is predicted to lose an additional 2.4 kilograms of weight.

Whereas there is a moderately strong positive linear relationship between weight loss and adherence level, we cannot draw a causal link between these two variables. Even though the study was a randomized comparative experiment, the variable imposed by the researchers was the diet plan, not the adherence level. Therefore, for the purpose of relating adherence level and weight loss, this study is essentially observational. However, we still might be interested in investigating whether the relationship observed in this sample is strong enough to convince us that it did not arise by chance. However, these subjects also were not a random sample from a larger population. (They volunteered for this study in response to advertisements.) Still, we might cautiously consider them representative of overweight men and women from the Northeast who would consider enrolling on these diets. With this consideration, we can proceed with a test to determine whether the observed level of association is higher than would be expected by random variation alone.

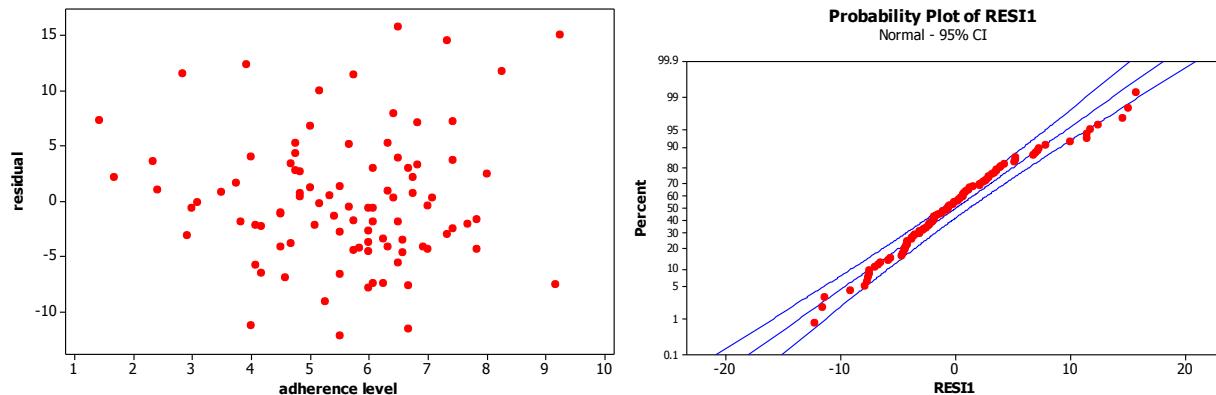
$H_0: \beta_1 = 0$ , where  $\beta_1$  represents the population slope coefficient. This null hypothesis indicates that there is no linear relationship between adherence level and amount of weight loss.

$H_a: \beta_1 \neq 0$  There is a linear relationship in the population.

Minitab produces the following output:

| Predictor       | Coef   | SE Coef | T     | P     |
|-----------------|--------|---------|-------|-------|
| Constant        | -8.432 | 2.306   | -3.66 | 0.000 |
| adherence level | 2.3876 | 0.3971  | 6.01  | 0.000 |

Checking the other technical conditions for the regression model, we find the following residual plots:



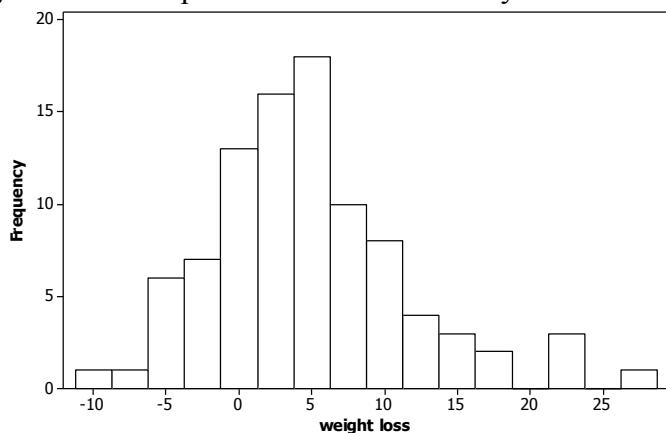
The plot of residuals vs. the explanatory variable does not reveal any serious problems, although there is a (slight?) suggestion of increasing variability with larger values. The normal probability plots suggests a bit of a skew to the right but not too much. This last condition is a bit less problematic with the large sample size in this study. We could consider a transformation to account for the increasing variability, but the increase does not seem substantial enough to warrant a transformation here. These technical conditions seem to be fairly well met.

The test statistic is very large ( $t = 6.01$ ) and the p-value very small (0.000 to three decimal places), and so we can conclude that the experimental data provide extremely strong evidence that there is an association between adherence level and weight loss. The p-value reveals that it would be almost impossible to obtain such large sample correlation and slope coefficients from random sampling variation alone. At any reasonable significance level, we conclude that this association is statistically significant.

We can follow up with a 95% confidence interval for the population slope  $\beta_1$ . We obtain  $2.388 \pm 1.986(0.397)$ , which is  $2.388 \pm 0.788$ , or  $(1.600, 3.176)$ . This interval suggests that the additional predicted weight loss for each additional step of adherence to diet is between 1.6 and 3.2 kilograms.

But remember the caveat that we mentioned earlier: the subjects' adherence levels were observed and not imposed, so we cannot draw a cause-and-effect conclusion between adherence level and weight loss. Furthermore, we must be cautious in stating to what population we are willing to generalize these conclusions.

(d) Mean weight loss: To address the question of whether dieters who complete 12 months on one of these popular diet plans actually tend to lose weight, we can begin by combining the weight loss diet across all four diet plans. This step seems reasonable because of our conclusion in (a) that there is no evidence of an effect of diet plan on weight loss. Although the adherence level does appear to be related to amount of weight lost, the completion rate did not vary significantly across the diets, providing further justification for pooling across the diets. A histogram of the weight loss amounts (in kilograms) for the 93 subjects who completed the 12-month study follows:



This histogram reveals that the distribution of weight loss amounts is a bit skewed to the right. The mean weight loss is 4.95 kilograms, with a standard deviation of 6.89 kilograms. The median is 3.90 kilograms, and most of the subjects had a positive weight loss; in fact, 71 of 93 (76.3%) did.

To perform statistical inference with these data, we need to again consider the volunteer nature of the sample and that any randomness here is hypothetical. We will proceed to conduct tests and make inferences, which will tell us whether the sample results are extreme enough to be unlikely to occur by random variation alone, but we need to keep in mind that the sample may not be representative of any population.

Even though the distribution is a bit skewed, the large sample size ( $n = 93$ ) allows us to perform a  $t$ -test of the hypotheses:

$H_0: \mu = 0$  (the mean weight loss in the population of dieters who could use one of these popular plans is zero)

$H_a: \mu > 0$  (the mean weight loss in the population of dieters who could use one of these popular plans is positive)

The test statistic turns out to be to  $t_0 = \frac{4.95 - 0}{6.89/\sqrt{93}} = 6.92$ , producing a  $p$ -value of essentially zero. This

suggests that the sample data provide overwhelming evidence that the population mean weight loss exceeds zero, that is, that dieters on these plans do tend to lose weight on average. A 95% confidence interval for  $\mu$  turns out to be (3.53, 6.36), so we can be 95% confident that the population mean weight loss is between 3.53 and 6.36 kilograms.

We can cautiously follow this up with a 95% prediction interval for the weight loss of an individual

dieter:  $4.95 \pm 1.986(6.89)\sqrt{1 + \frac{1}{93}}$ , which is  $4.95 \pm 13.76$ , which is (-8.81, 18.71). This interval implies

that, with 95% confidence, we can assert only that an individual dieter is predicted to see a weight change anywhere between a gain of 8.8 kilograms and a loss of 18.7 kilograms. However, the slight skewness in the sample data leads us to question the validity of this prediction interval because the normality condition is essential for this procedure.

We could also perform a “sign test” of the hypotheses. A sign test reduces the analysis to counting how many “positive” and “negative” differences there are and determining whether we have significantly more of one than the other.

$H_0: \pi = 0.5$  (half of the population of all potential dieters would lose positive weight on one of these diet plans)

$H_a: \pi > 0.5$  (more than half of the population of all potential dieters would lose positive weight on one of these diet plans)

The data reveal that 71 of 93 subjects had positive weight loss. The binomial distribution (with parameters  $n = 93$  and  $\pi = 0.5$ ) reveals that the  $p$ -value of  $P(X \geq 71)$  equals essentially zero ( $z \approx 5.08$ ). Thus, this sign test leads to a similar conclusion: overwhelming evidence that more than half of the population would lose positive weight.

Because of the volunteer nature of the sample, it is not completely clear to what population we can generalize these results. Moreover, even though we concluded that the mean weight loss is significantly larger than zero, we cannot attribute the cause to the diet. Without the use of a comparison group of people who did not participate in a diet plan, we cannot conclude that the diet alone is responsible for the tendency to lose weight. Perhaps even the power of suggestion from being in the study was a sufficient cause for these individuals to lose weight on average.

Summarizing our findings from this study:

- We do not have convincing evidence to suggest that one of these popular diet plans produces more weight loss on average than another.
- We do not have convincing evidence to suggest that completion rates differ among the four diet plans.
- We do have very strong evidence to suggest that dieters on one of these plans will lose weight, between 3.5 and 6.5 kg on average, subject to the caveat that the dieters in this study may not be representative of a larger population of overweight people.
- We do have significant evidence that those who adhere to a diet more closely do tend to lose more weight (subject to the same caveat).

## CHAPTER 5 SUMMARY

In this chapter, you explored procedures for analyzing two or more groups and for analyzing relationships. As in earlier chapters, you saw that different study designs could lead to the same mathematical computations, but also remember to consider the study design when drawing your final conclusions. In particular, you should think about which variables were controlled by the researchers and which were not.

A very common procedure to use with two-way tables is the chi-square procedure. You saw that this reduces to the two-sample  $z$ -test in the case of two groups. The chi-square procedure is a large sample procedure and Fisher's Exact Test can be used with smaller sample sizes. A parallel procedure for comparing several means is Analysis of Variance, which models the ratio of the variability between groups to the variability within groups using the  $F$  distribution.

Regression is a very important field of study and your next course in statistics will probably be entirely about regression models. We scratched the surface here by looking at the appropriate numerical and graphical summaries for analyzing the relationship between two quantitative variables (scatterplots and correlation coefficient) and least squares regression models for linear relationships. In deciding whether you have a statistically significant relationship, the  $t$ -statistic uses the common test statistic form that you saw in earlier chapters (observed – hypothesized)/standard error, where the standard error of the sample slopes depends on the sample size, the variability in the  $x$  values and the vertical spread about the regression line.

These calculations are complex to carry out by hand, but our focus is on knowing when to use which procedure, how to check the validity of the procedure, and how to interpret the resulting output. Learning to effectively interpret and communicate statistical results is at least as important a skill as learning which computer menus to use!

## TECHNOLOGY SUMMARY

- In this chapter, you learned how to use Minitab/R to perform chi-square tests, ANOVA, and inference for regression. You also learned how to create scatterplots, including coded scatterplots, and calculate correlation coefficients.
- You used applets to explore properties of  $F$ -statistics, regression lines, and regression coefficients. Though we hope these have given you some memorable visual images, you will generally use R or Minitab to carry out specific analyses.

## SUMMARY OF PROCEDURES FOR COMPARING SEVERAL POPULATIONS, EXPLORING RELATIONSHIPS

| Setting                                | EV: Categorical<br>RV: Categorical                                                                                                                                                                                                                 | EV: Categorical<br>RV: Quantitative                                                                                                                                                                                     | EV: Quantitative<br>RV: Quantitative                                                                                                                                                                                                                                                                                                                                           |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Graphical summary                      | Segmented bar graph                                                                                                                                                                                                                                | Stacked boxplots                                                                                                                                                                                                        | Scatterplot                                                                                                                                                                                                                                                                                                                                                                    |
| Numerical summary                      | Conditional proportions                                                                                                                                                                                                                            | Group means and standard deviations                                                                                                                                                                                     | Correlation coefficient                                                                                                                                                                                                                                                                                                                                                        |
| Procedure name                         | Chi-square                                                                                                                                                                                                                                         | ANOVA                                                                                                                                                                                                                   | Regression                                                                                                                                                                                                                                                                                                                                                                     |
| Valid to use theory-based procedure if | <ul style="list-style-type: none"> <li>All expected counts <math>\geq 1</math>, at least 80% of expected counts <math>\geq 5</math></li> <li>Data are simple random samples or independently chosen random samples or random assignment</li> </ul> | <ul style="list-style-type: none"> <li>Independent SRSs or random assignment</li> <li>Populations normal (probability plots of samples)</li> <li>Variances are equal (<math>s_{\max}/s_{\min} &lt; 2</math>)</li> </ul> | <ul style="list-style-type: none"> <li>Linear relationship (residuals vs. <math>x</math>)</li> <li>Independent observations (e.g., random samples or randomization)</li> <li>Normality of response <i>at each x value</i> (probability plot/histogram of residuals)</li> <li>Equal variance of <math>y</math> <i>at each x value</i> (residuals vs. <math>x</math>)</li> </ul> |
| Null hypothesis                        | $H_0$ : Response variable distributions are the same or $H_0$ : no association between response variable and explanatory variable                                                                                                                  | $H_0: \mu_1 = \dots = \mu_I$ or $H_a$ : treatment means are equal                                                                                                                                                       | $H_0: \beta_1 = 0$ (no relationship between response variable and explanatory variable)                                                                                                                                                                                                                                                                                        |
| Test statistic                         | $X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$                                                                                                                                                                                              | $F = MST/MSE$                                                                                                                                                                                                           | $t_0 = \frac{b_1 - hypothesized\ value}{SE(b_1)}$                                                                                                                                                                                                                                                                                                                              |
| Distribution for p-value               | Chi-squared with $(c - 1)(r - 1)$ df                                                                                                                                                                                                               | $F$ with $I - 1, N - I$ df                                                                                                                                                                                              | $t$ with $n - 2$ df                                                                                                                                                                                                                                                                                                                                                            |

## Quick Reference to ISCAM R Workspace Functions and Other R Commands

| Procedure Desired                          | Function Name (Options)                                                                                                 |
|--------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| Normal Probability Plot                    | qqnorm(data)                                                                                                            |
| Probability Plot                           | qqplot(data from distribution, your data)                                                                               |
| Probabilities from Chi-Square distribution | iscamchisqprob(xval, df)<br>returns upper tail probability                                                              |
| Chi-square Test                            | chisq.test(table(data))<br>chisq.test(table(data))\$expected \$residuals                                                |
| One-way ANOVA                              | summary(aov(response~explanatory))                                                                                      |
| Probabilities from F distribution          | pf(x, df1, df2, lower.tail=FALSE)                                                                                       |
| Scatterplot                                | plot(explanatory, response)                                                                                             |
| Correlation coefficient                    | cor(x, y)                                                                                                               |
| Least Squares Regression Line              | lm(response~explanatory)<br>lm(resonse~explanatory)\$residuals                                                          |
| Coded Scatterplot                          | plot(response~explanatory, col=groups)                                                                                  |
| Inference for regression                   | summary(lm(response~explanatory))                                                                                       |
| Prediction intervals                       | predict(lm(response~explanatory),<br>newdata=data.frame(explanatory-value),<br>interval = "prediction" or "confidence") |
| Superimpose $y = x$ line                   | lines(response, response)                                                                                               |
| Superimpose regression line                | abline(lm(response~explanatory))                                                                                        |

## Quick Reference to Minitab Commands

| Procedure Desired                          | Menu                                                                                               |
|--------------------------------------------|----------------------------------------------------------------------------------------------------|
| Probability Plot                           | Graph > Probability Plot                                                                           |
| Probabilities from Chi-Square distribution | Graph > Probability Distribution Plot                                                              |
| Chi-square Test                            | Stat > Tables > Chi-Square Test for Association<br>Stat > Tables > Cross Tabulation and Chi-Square |
| One-way ANOVA                              | Stat > ANOVA > One-way                                                                             |
| Probabilities from F distribution          | Graph > Probability Distribution Plot                                                              |
| Scatterplot                                | Graph > Scatterplot                                                                                |
| Correlation coefficient                    | Stat > Basic Statistics > Correlation                                                              |
| Least Squares Regression Line              | Stat > Regression > Fitted Line Plot<br>Storage: Residuals                                         |
| Coded Scatterplot                          | Graph > Scatterplot, With Groups                                                                   |
| Inference for Regression                   | Stat > Regression > Regression > Fit Regression Model                                              |
| Prediction Intervals                       | After running the model:<br>Stat > Regression > Regression > Predict                               |
| Superimpose $y = x$ line                   | right click, Add > Calculated Line                                                                 |
| Superimpose regression line                | Stat > Regression > Fitted Line Plot<br>right click on scatterplot, Add > Regression Fit           |

**INDEX**

- Addition rule, 16
- Adjusted Wald interval, 87
- Alternative hypothesis, 35
- ANOVA, 335
- Applet
  - Analyzing two quantitative variables, 360, 381
  - Analyzing two-way tables, 212, 324
  - Comparing groups (quantitative), 270, 334, 339
  - Dolphin study, 208
  - Guess the Correlation, 355
  - Matched pairs randomization, 287
  - Normal probability calculator, 70, 80
  - One proportion inference, 23, 39
  - Power simulation, 53
  - Random Babies, 13
  - Randomizing subjects, 199
  - Reese's Pieces, 63
  - Sampling from finite population, 147
  - Sampling words, 101
  - Simulating ANOVA tables, 342
  - Simulating confidence intervals, 83, 90
  - Theory-Based inference, 76
- Bar graph, 20, 32
- Baseline rate, 221
- Basic regression model, 384
- Biased, 94
- Binary, 34
- Binomial distribution, 27
- Binomial intervals, 48
- Binomial power, 60
- Binomial probabilities, 32
- Binomial rejection region, 60
- Binomial test, 36
- Blinding (experiment), 203
- Boxplot, 140
- Case-control study, 228
- Categorical, 34
- Center of distribution, 134
- Central limit theorem
  - sample proportion, 66, 103
- Central Limit Theorem
  - difference in proportions, 186
  - Chi-square distribution, 317
  - Chi-square test statistic, 315
  - Coefficient of determination, 364
  - Cohort study, 228
  - Comparative experiment, 203
  - Comparing two proportions, 190
  - Complement rule, 16
  - Conditional proportions, 182
  - Confidence interval, 47
  - Confidence level, 84
  - Confounding variables, 193
  - Continuity correction, 75
  - Convenience sample, 94
  - Correlation coefficient, 351
  - Coverage rate, 84
  - Critical value, 81
  - Cross-classified study, 228
  - Cross-sectional design, 228
  - Distribution, 134
  - Double blind, 203
  - Duality, 49
  - Exact binomial inference, 51
  - Expected value, 16, 41
  - Experimental study, 197
  - Experimental units, 197
  - Explanatory variable, 191
  - Extrapolation, 364
  - F* distribution, 335
  - F* statistic, 335
  - Finite population correction factor, 109
  - Fisher's exact test, 214
  - Hawthorne effect, 203
  - Homogeneity of proportions, 319
  - Hypergeometric distribution, 212, 213
  - hypergeometric random variable, 108
  - Hypotheses, 35
  - Hypothesized value, 47
  - Independence, 28
  - Influential observation, 372
  - Intercept coefficient, 363
  - Interquartile range, 139
  - Interquartile range (IQR), 140

- Least squares line, 361  
 Least squares regression line, 361  
 Level of significance, 53  
 Linear association, 348  
 Long-run proportion, 13  
 Margin-of-error, 81  
 Matched-pairs design, 281  
 Mean of absolute differences (MAD), 338  
 Mutually exclusive, 28  
 Negative association, 348  
 Non-sampling errors, 105  
 Normal approximation to binomial, 76  
 Normal probability calculations, 69  
 Normal probability curve, 65  
 Null distribution, 21  
 Null hypothesis, 35  
 Null model, 21  
 Observational study, 197  
 Observational units, 34  
 Odds, 228  
 Odds ratio, inference, 232  
 One proportion  $z$ -intervals, 82  
 One proportion  $z$ -procedures, 112  
 One proportion  $z$ -test, 78  
 One sample  $t$ -test, 159  
 One sample  $z$ -interval, 81  
 One-sided alternative, 43  
 Outlier, 140  
 Paired data, 283  
 Paired  $t$ -procedures, 293  
 Paired  $t$ -test, 288  
 Parallel dotplots, 133, 252, 283  
 Parallel histograms, 134  
 Parameter, 34, 93  
 Percentage change, 188  
 Percentiles, 80  
 Placebo effect, 203  
 Plausible, 26  
 Population, 92  
 Positive association, 348  
 Power, 54  
 Practical significance, 114  
 Prediction interval, 162  
 Probability, 13, 24  
 Probability rules, 16  
 Prospective design, 228  
 p-value, 24  
 Quantitative, 34  
 Random assignment, 198  
 Random variable, 15  
 Randomization test, 270, 271  
 Randomized experiment, 203  
 Regression line, 361  
 Rejection region, 53  
 Relative frequency, 13  
 Relative risk, 221  
 Relative risk, inference, 226  
 Residual, 325, 336, 358  
 Residual plots, 385  
 Response variable, 191  
 Retrospective design, 228  
 Sample, 92  
 Sample size, 34  
 Sampling frame, 95  
 Sampling variability, 94  
 Scatterplot, 347  
 Scope of conclusions, 25  
 Segmented bar graphs, 182  
 Shape of distribution, 134  
 Sign test, 413  
 Significance, 24  
 Simple random sample, 94  
 Simple random sample, selecting, 95, 253  
 Skewed distributions, 134  
 Skewed to the left, 134  
 Skewed to the right, 134  
 Slope coefficient, 363  
 Spread. *See Variability*  
 Squared residuals, 359, 361  
 Standard deviation, normal curve, 65  
 Standard deviation, of random variable, 17  
 Standard deviation, sample proportion, 65  
 Standard error  
 difference in means, 272  
 difference in proportions, 187  
 sample mean, 154  
 Standard error (sample proportion), 79  
 Standard normal distribution, 80  
 Standard score, 69  
 Statistic, 26, 34  
 Statistically significant, 24  
 Stratified random sample, 130

- Strength of association, 348  
Sum of absolute errors (SAE), 361, 367  
Sum of squared residuals, 361  
Systematic random sampling, 292  
Tally, 32  
Test statistic, 73  
Transformation, 141  
Treatment, 197  
Two sample  $z$ -procedures, 189  
Two-sample  $t$ -test, 258  
Two-sample  $z$ -test, 187  
Two-sided alternative, 43  
Two-sided p-value, 44  
Two-way table, 181  
Type I error, 54  
Unbiased, 94  
Variability of distribution, 134  
Variable, 34  
Variance, of random variable, 17  
Wald interval, 81  
Wilson adjustment, 87, 188