Elementary Statistics

MAT118

OER Book, 3rd Edition, Fall 2022

Dutchess Community College
Math & Computer Science Department
Chapter 1 Introduction To Statistics and Simulation
* Section 1.1 Introduction to Statistics and Key Terms ....................................................... 2
* Section 1.2 Introduction to Distribution of Sample Statistics ............................................. 16
* Section 1.3 Understanding when Statistics are Unusual ..................................................... 30
* Section 1.4 Introduction to Computer Simulation ............................................................. 46
* Section 1.5 Additional Details on Quantitative Data ....................................................... 62

Chapter 2 Introduction To Hypothesis Testing
* Section 2.1 Introduction to Hypothesis Testing ................................................................. 75
* Section 2.2 Introduction to P-value and Conclusions in a Hypothesis Test ......................... 94
* Section 2.3 Standardized Statistics and Conclusions in a Hypothesis Test ....................... 112
* Section 2.4 Strength of Evidence in a Conclusion ............................................................ 135

Chapter 3 Theory-based Statistics and Comparing Means and Proportions
* Section 3.1 Sampling ........................................................................................................ 146
* Section 3.2 Association vs. Causation ............................................................................. 156
* Section 3.3 Theory-Based Approach to Inference ........................................................... 168
* Section 3.4 Comparing Two Means ................................................................................ 181
* Section 3.5 Comparing Two Proportions ........................................................................ 193

Chapter 4 Confidence Intervals
* Section 4.1 Introduction to Confidence Intervals .............................................................. 204
* Section 4.2 Theory-Based Approach to Confidence Intervals .......................................... 216
* Section 4.3 Impacts on the Width of a Confidence Interval .............................................. 224

Chapter 5 Correlation and Regression
* Section 5.1 Scatterplots and Review of Linear Functions ................................................. 230
* Section 5.2 The Regression Equation ............................................................................... 244
* Section 5.3 Prediction and Residuals ................................................................................. 254
* Section 5.4 Testing for Correlation ................................................................................. 268
Chapter 1 | Introduction to Statistics and Simulation

Contents of Chapter 1

Section 1.1 Introduction to Statistics and Key Terms

Section 1.2 Introduction to the Distribution of Sample Statistics

Section 1.3 Understanding When Statistics are Unusual and Why We Care

Section 1.4 Introduction to Computer Simulation

Section 1.5 Additional Details on Quantitative Data

You are probably asking yourself the question, "When and where will I use statistics?" If you watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you watch a television news program or see statistics on the Internet, you are given sample information. This means that you are given information about a subset of a larger group. The subset group is called the sample, and the larger group is called the population. With this information, you may make a decision about the correctness of a statement, claim, or "fact" related to the entire population. Statistical methods can help you make the "best educated guess" about some aspect of the population.

Since you will undoubtedly be given statistical information, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, and early childhood development require at least one course in statistics. Included in this chapter are the basic ideas and vocabulary of probability and statistics. You will soon understand that probability and statistics work together. You will also learn how data are gathered and that "good" data can be distinguished from "bad" data.

Organizing and summarizing data is called descriptive statistics. Two ways to summarize data are by graphing and by using numbers (for example, finding an average).

In this course, we focus our time on learning formal methods for drawing conclusions from "good" data. The formal methods are called inferential statistics. Statistical inference uses probability to determine how confident we can be that our conclusions are correct. Effective interpretation of data (inference) is based on appropriate procedures for producing data and thoughtful examination of the data.

The goal of statistics is to gain an understanding of sample data in order to make conclusions about some aspect of the population. The calculations can be done using technology. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.
### Key Terms

**Observational Units**
The individual entities on which data are recorded. For example, I observe _cars_ and record the _make_ (Honda or Ford or Toyota) and _mileage_ (25,000 or 150,000 or 75,000) of each car.

**Variables**
The characteristics of the observational units that are being recorded. For example, I observe _cars_ and record the _make_ (Honda or Ford or Toyota) and _mileage_ (25,000 or 150,000 or 75,000) of each car.

**Data**
The recorded results from a study. For example, in a study about cars the collected data is a list of makes and mileages (Honda 25,000, Ford 150,000, and Toyota 75,000).

**Quantitative Variable**
Variable whose outcomes are a set of numbers. In a study about cars, _mileage_ is a quantitative variable.

**Categorical Variable**
Variable whose outcomes are a set of categories. In a study about cars, _make_ is a categorical variable.
**Population**
The entire collection of persons, things, or objects that we want to learn about. When describing a population, we often use the word “all”. For example, in a study about cars, the population is all cars sold in the United States since 2010.

**Sample**
A sample is a portion (or subset) of the population. We study that portion (the sample) to gain information about the population. The sample is made up of observational units.

---

**Parameter**
A numerical characteristic of a population. Its value is usually unknown because we typically can’t gather information from every single element of a population. In general, the purpose of our statistical study is to try to understand something about the value of the unknown population parameter. For example, in a study about cars, we may want to estimate the proportion of all cars sold in the United States since 2010 that are Toyotas. Or, we may want to estimate the average mileage of all cars sold in the United States since 2010.

**Statistic**
A numerical characteristic of an observed sample. The value of the statistic can be calculated from the collected sample data. Therefore, the value of the statistic is a known, observed value that we get from our sample, and we then use the statistic to help us understand and make conclusions about the unknown value of the population parameter. For example, in a study about cars, we may find that in a sample of 100 cars sold in the United States since 2010, 43% are Toyotas, and that the average mileage of these 100 cars is 83,295 miles.

\[ \pi \] (pi)
This is the symbol for the parameter when the variable is categorical. It can be described as the proportion of observational units in the population that belong to a particular category, often referred to as the “success” category. For example, in a study about cars, \( \pi \) (or pi) is the proportion of all cars sold in the United States since 2010 that are Toyotas. The symbol \( \pi \) is pronounced “pie”. Alternatively, \( \pi \) can be described as a long-run proportion of successes. Its value is typically unknown.

\[ \hat{p} \] (p-hat)
This is the symbol for the statistic when the variable is categorical. It can be described as the proportion of observational units in the sample that belong to the success category and its value can be calculated from the observed data. For example, in a study about cars, \( \hat{p} \) (or p-hat) is the proportion of Toyotas in a sample of 100 cars sold in the United States since 2010, and its value is \( \hat{p} = 0.43 \) (or 43%). The symbol \( \hat{p} \) is pronounced “p-hat”.

\[ \mu \] (mu)
This is the symbol for the parameter when the variable is quantitative. It is the mean of the values of the variable for the entire population. For example, in a study about cars, \( \mu \) (or mu) is the mean mileage of all cars sold in the United States since 2010. The symbol \( \mu \) is pronounced “mew”. Its value is typically unknown.
\( \bar{x} \) (x-bar)
This is the symbol for the statistic when the variable is quantitative. It can be described as the mean of the values of the variable from the sample and its value can be calculated from the observed data. For example, in a study about cars, \( \bar{x} \) (or x-bar) is the mean mileage from a sample of 100 cars sold in the United States since 2010, and its value is \( \bar{x} = 83,295 \). The symbol \( \bar{x} \) is pronounced “x-bar”.

\( n \)
\( n \) is the symbol for the sample size. This is the number of observational units in the sample.

Probability of an event
The proportion of times that the event occurs in a long run of trials under identical conditions.

<table>
<thead>
<tr>
<th>Example 1.1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine what the key terms and symbols refer to in the following study. We want to know the average (mean) amount of money that all first-year college students spent on school supplies at ABC College this term. We randomly survey 100 first-year students at the college. Three of those students spent $150, $200, and $225, respectively. We find that our sample of 100 students had an average of $218.74 spent on school supplies.</td>
</tr>
<tr>
<td>Solution 1.1.1</td>
</tr>
<tr>
<td>The observational units are first-year students at ABC College.</td>
</tr>
<tr>
<td>The variable is the amount of money spent on supplies by each first-year student. This is a quantitative variable.</td>
</tr>
<tr>
<td>The population is all first-year students attending ABC College this term.</td>
</tr>
<tr>
<td>The sample (the set of observational units) is the 100 first-year students at ABC College that were surveyed.</td>
</tr>
<tr>
<td>The sample size is ( n = 100 ).</td>
</tr>
<tr>
<td>The parameter is the average (mean) amount of money spent on school supplies by all first-year college students at ABC College this term. The symbol for this parameter is ( \mu ) and its value is unknown.</td>
</tr>
<tr>
<td>The statistic is the average (mean) amount of money spent on school supplies this term by the 100 surveyed first-year students at ABC College. The symbol for this statistic is ( \bar{x} ) and its value is ( \bar{x} = 218.74 ).</td>
</tr>
<tr>
<td>The data are the dollar amounts spent by the 100 first-year students in the sample. Examples of three data values are $150, $200, and $225.</td>
</tr>
</tbody>
</table>
1.1.1
Determine what the key terms and symbols refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by all families with children at Knoll Academy. We randomly survey 85 families with children in the school. Three of the families spent $65, $75, and $95, respectively. We find that our sample of 85 families spent an average (mean) of $72.09.

Example 1.1.2
Determine what the key terms and symbols refer to in the following study. A study was conducted at a local college to analyze the average cumulative GPAs of students who graduated last year. Fill in each blank with the letter of the phrase that best describes the item.


a) the cumulative GPA of each student who graduated from the college last year
b) 3.65, 2.80, 1.50, 3.90
c) a group of selected students who graduated from the college last year
d) the average cumulative GPA of all students who graduated from the college last year
e) all students who graduated from the college last year
f) the average cumulative GPA of the students in the study who graduated from the college last year

Solution 1.1.2
1. e 2. f 3. d 4. c 5. a 6. b
Example 1.1.3

Determine what the key terms and symbols refer to in the following study. As part of a study designed to test the safety of automobiles, researchers at the National Transportation Safety Board (NTSB) collected and reviewed data about the effects of automobile crashes on test dummies. Following are the criteria they used:

<table>
<thead>
<tr>
<th>Speed at which cars crashed</th>
<th>Location of dummy</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 miles per hour</td>
<td>Front seat</td>
</tr>
</tbody>
</table>

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour (mph). The NTSB wants to know the proportion of all crashes in which the dummy in the driver’s seat would have had head injuries if it had been an actual driver. They start with a simple random sample of 75 crashes and find that in 15 of them the driver would have had head injuries.

Solution 1.1.3

The **observational units** are crashes at 35 mph.

The **variable** is whether the dummies would have had a head injury if it had been an actual driver. This is a **categorical variable**.

The **population** is *all* crashes at 35 mph containing dummies in the front seat.

The **sample** is the 75 crashes the researchers observed.

The **sample size** is *n*=75.

The **parameter** is the proportion of *all* crashes in which the dummy (if it had been an actual driver) would have suffered head injuries. The value is an unknown proportion which the researchers are trying to understand. The symbol for this parameter is *π*.

The **statistic** is the proportion of the sampled 75 crashes in which the dummy would have suffered head injuries if it had been an actual driver. Its value is the known observed proportion, 15/75 = 0.2 = 20%, that can be calculated from the sample data. The symbol for the statistic is *p̂* and its value is *p̂* = 0.2.

The **data** are either “yes, had head injury,” or “no, did not have head injury.”
Example 1.1.4

Determine what the key terms and symbols refer to in the following study. An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit is 160 of the doctors.

Solution 1.1.4

The **observational units** are medical doctors.

The **variable** is whether the medical doctor has been involved in one or more malpractice suits. This is a **categorical variable**.

The **population** is *all* medical doctors listed in the professional directory.

The **parameter** is the proportion of *all* medical doctors who have been involved in one or more malpractice suits. The value is an unknown proportion which the insurance company is trying to understand. The symbol for this parameter is \( \pi \).

The **sample** is the 500 medical doctors selected at random from the professional directory.

The **sample size** is \( n = 500 \).

The **statistic** is the proportion of the 500 medical doctors in the sample who have been involved in one or more malpractice suits. Its value is the known, observed proportion, \( \frac{160}{500} = 0.32 = 32\% \), that can be calculated from the sample data. The symbol for the statistic is \( \hat{p} \) and its value is \( \hat{p} = 0.32 \).

The **data** are either “yes, was involved in one or more malpractice lawsuits,” or “no, was not involved in any malpractice lawsuits.”
Probability is the Long Run Proportion of Success

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the result may not be two heads and two tails. However, if you toss the same coin 4,000 times, it is very likely that the result will be very close to half heads and half tails. The expected theoretical probability of heads in any one toss is \( \frac{1}{2} = 0.5 \). Even though the outcome of a few tosses is uncertain, the outcome is more predictable when there are a large number of tosses.

A person tossed a coin 2,000 times and the results were 996 heads. The fraction 996/2000 is equal to 0.498, which is very close to the expected probability of 0.5. In general, the probability of an event is the long-run proportion of times that the event occurs. We would say that the probability of the coin landing on heads is \( \frac{1}{2} = 0.5 \) because in the long run (infinitely many tosses!) it will land on heads \( \frac{1}{2} = 50% \) of the time.

The probability of an event is the proportion of times that the event occurs in a large number of trials.
The class should break into groups of about 3 students each. The instructor will provide each group with a deck of cards. Students should be aware that the teacher may have removed and/or added cards to the deck so that it may no longer be a “fair” deck of cards that you are used to seeing.

Group members should NOT view the faces of the cards. The purpose of this exercise is to approximate the proportion of cards in the deck without actually counting the entire deck.

Each group should choose a “dealer,” a “player”, and a “recorder.” If possible, the dealer should be a person who is able to successfully shuffle cards.

(a) The dealer should shuffle the deck of cards provided by the teacher. The dealer should then allow the player to select 6 cards of their choice.

(b) The group should now view the 6 selected cards and record the number of cards (out of 6 total) which are red cards. The recorder should create a table similar to that shown below and record the number of red cards found in this first trial in the appropriate box.

<table>
<thead>
<tr>
<th>Trial #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># red</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cards</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(c) The 6 selected cards should be added back into the deck, and the group should now repeat steps (a) and (b) in order to re-shuffle, select a new set of 6 cards, and record the number of red cards selected. The group should repeat this process and record the number of red cards for a total of 20 trials. Be sure to shuffle the cards well between each trial.

(d) The group should now create a dot plot to accompany the table from part (b). Be sure to label the meaning and values along the horizontal axis. Also use a complete sentence to identify the meaning of one dot in the dot plot.

(e) Add another row to the table, “proportion of red cards”, and for each trial, write the proportion of cards that were red. For example, if a trial had 2 cards that were red out of a total of 6 cards, the proportion would be written as $2/6 = 0.33$. Fill in the bottom row by writing each value in the “# red cards” row as a proportion.

(f) The group should now create a second dot plot based on the new row from part (e). Be sure to label the meaning and values along the horizontal axis. Also identify the meaning of each dot in the dot plot.

(g) As a group, use the dot plot to identify the group’s BEST ESTIMATE of the proportion of ALL THE CARDS IN THE DECK that are red. Explain how the dot plot helped the group come to this conclusion.

(h) As a group, decide how much confidence the group has in the estimate from part (g). Also describe how the group could go about becoming more confident in their estimate (other than counting every single card in the deck).

(i) Describe the meaning of the parameter in this situation. Give the symbol for the parameter. Is the value of the parameter known or unknown?
Sam wants to determine whether his die is fair. He decides to roll the die 12 times and record the number of times that the die lands on the number two. He records the results in the table below where “y” represents “yes, a two was rolled” and “n” represents “no, a two was not rolled”.

<table>
<thead>
<tr>
<th>Trial #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rolled a two?</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>

(a) Describe the parameter of interest in context. What is the symbol for this parameter? Is the value of the parameter known or unknown?

(b) Identify (describe in words) the statistic from Sam’s sample. What is the symbol for this statistic? Is the value of the statistic known or unknown?

(c) Using the collected sample, what is Sam’s best estimate for the probability of rolling the number two in the long run? In other words, what is the best estimate for the value of the population parameter?

(d) Describe how much confidence you have in the estimate from part (c). Also describe how you could become more confident in the estimate.

(e) Based on this sample, do you think that this die is fair? Explain. Also explain how you could become more confident in your conclusion about the fairness of the die.

(f) Suppose Sam now brings this die to work and asks many friends, acquaintances, and customers to help him with his little experiment. He manages to get about 200 people to each roll the die 12 times and record the proportion of times the die rolls a “two”. A dot plot showing the sample statistic from each person is shown below. Given this dot plot, what are your new, updated answers for questions (c), (d), and (e) above? Explain.

Solution 1.1.6

(a) The parameter of interest is the long run proportion of times that the die rolls the number two. The symbol for the parameter is $\pi$, and its value is unknown.

(b) The statistic from this sample is the number of twos rolled divided by the total number of rolls. The value is a known value that we observed from the sample, and it has symbol $\hat{p}$. So, the value of the statistic is $\hat{p} = \frac{5}{12} \approx 0.417 = 41.7\%$.

(c) Sam’s best estimate of the probability of this die rolling a two is about 41.7% of the time since that is what he observed in the sample. The value 41.7% is the best guess for the value of the population parameter.
Solution 1.1.6 continued

(d) We don’t have a great deal of confidence in Sam’s estimate that the die rolls the number two 41.7% of the time in the long run. Sam only rolled the die 12 times. He could be more confident if the die were rolled more times, or if he had a better sense of how usual/unusual it is to get the result he actually observed. Is a result of $5/12 = 41.7\%$ twos a usual sort of statistic to get when rolling the die 12 times, or was it just a fluke?

(e) When rolling a fair die 12 times, we would expect to get about $1/6 = 0.167 = 16.7\%$ of the rolls to land on the number two. That means that when rolling a fair die 12 times, we would expect that about 2 of the rolls to land on the number two. Also notice that it would be usual and expected if we rolled the number two approximately 2 times rather than expecting exactly 2 times. Getting 1 roll out of 12 landing on the number two or getting 3 rolls out of 12 landing on the number two, for example, would be usual and expected if the die were fair. Sam’s sample rolled a two 5 times out of 12, which does seem like an unusually high number of times to roll the number two if the die were fair. So, we perhaps have a die that is not fair. We don’t have much confidence in this conclusion since we only rolled the die 12 times. We could be more confident if Sam rolled the die a larger number of times.

(f) The dot plot is centered at around $0.33 = 33\%$ of the time rolling a two. Most people who participated in rolling the die 12 times got a sample statistic between 17% twos and 58% twos. So, our best guess is that it rolls the number two 33% of the time. A fair die would roll the number two $1/6 = 17\%$ of the time in the long run. It seems more likely that this die is rolling the number two more than $1/6$ of the time in the long run. Increasing the number of rolls (the sample size), and also gathering more samples (increasing the number of repetitions), would help us increase our confidence in our conclusion about the long run proportion.
Section 1.1 | Introduction to Statistics and Key Terms

Exercises

Questions 1-3 refer to the following scenario:
Many studies have investigated whether people tend to call “heads” or “tails” when they are asked to predict the outcome of a coin flip. To research this question we collected a random sample of 832 people, of whom 517 selected the choice “heads” when the coin was flipped.

1) (Multiple Choice) What are the observational units in the study?
   (a) The 517 people who chose “heads”.
   (b) The 315 people who chose “tails”.
   (c) The 832 people used in the study.
   (d) Whether the person chose “heads” or “tails”.

2) (Multiple Choice) What is the parameter of interest in the study?
   (a) The mean number of people out of a sample of 832 that choose “heads”.
   (b) The population proportion of people who choose “heads” during a coin flip.
   (c) The probability that people choose “heads” rather than “tails”.
   (d) Whether the average person will choose “heads” or “tails”.

3) (Multiple Choice) What symbol should be used to represent the sample statistic?
   (a) \(\pi\) (pi)
   (b) \(\hat{p}\) (p-hat)
   (c) \(\mu\) (mu)
   (d) \(\bar{x}\) (x-bar)
   (e) \(n\)

4) Victoria, a student at Poughkeepsie High School, stood at the east entrance and observed 106 students arriving and recorded whether they arrived by foot, bike, bus, or personal vehicle. They also recorded the number of minutes early/late each student arrived.
   (a) Identify (describe in words) the observational units in this study.
   (b) Identify (describe in words) one of the variables in this study and classify the variable as categorical or quantitative.
   (c) Identify (describe in words) the other variable in this study and classify the variable as categorical or quantitative.

5) Suppose that BudgetAir, a passenger airline, randomly chooses 8 of their passenger jets and decreases the legroom in each of the rows by 2 inches. The airline surveys the passengers after several flights (2,360 passengers in total) recording their satisfaction with the flight using the scale below. They also asked the question “On a scale of 1 to 10, 1 being not at all, and 10 being very definitely, how likely are you to fly with BudgetAir again”?

   (a) Identify (describe in words) the observational units and state the sample size (including the correct symbol) in this study.
   (b) Identify (describe in words) one of the variables in this study and classify the variable as categorical or quantitative.
   (c) Identify (describe in words) the other variable in this study and classify the variable as categorical or quantitative.
6) Determine whether each of the following variables is categorical or quantitative. Give two examples of data that might be obtained for each variable.
   (a) morning commute time in minutes
   (b) county of residence
   (c) preferred brand of toothpaste
   (d) method of transportation
   (e) pain level
   (f) vehicle condition
   (g) size of the dog
   (h) balloon radius in centimeters
   (i) zip code
   (j) mass of an object
   (k) student ID number
   (l) reading level

7) Determine whether each of the following descriptions would result in categorical or quantitative data.
   (a) We record the flavor of each observed food as either savory or sweet.
   (b) We record the amount of temperature change that occurred in degrees Celsius.
   (c) We record the width of each artery in mm.
   (d) We record whether the texture is smooth or rough.
   (e) We record whether the matter is liquid or solid or gas.
   (f) We record the wait time in minutes.
   (g) We record whether each person came in first place, second place, or third place.

8) Suppose that you are planning to collect data from a sample of prisons around the country. Note that the prisons (not the incarcerated individuals) are the observational units in this case.
   (a) State two quantitative variables that you could collect data about from each of the prisons.
   (b) State two categorical variables that you could collect data about from each of the prisons.

9) Suppose that the observational units in a statistical study are the patients arriving at the emergency room of a particular hospital during one month.
   (a) Identify two categorical variables that you could collect data about from each patient.
   (b) Identify two quantitative variables that you could collect data about from each patient.

10) Kelley Blue Book is a vehicle valuation and automotive research company that reports market value prices for new and used automobiles. These values are often used by auto dealerships to determine the selling prices of their vehicles.
    (a) Identify two categorical variables one might record about each vehicle to determine the value of the vehicle. For each variable, give two examples of data that might be obtained.
    (b) Identify two quantitative variables one might record about each vehicle to determine the value of the vehicle. For each variable, give two examples of data that might be obtained.
11) As an academic counselor at Dutchess Community College, you are tasked with identifying students who are “at-risk” of being placed on double-secret academic probation. If students are deemed “at-risk” then an early intervention counseling session can be scheduled to identify issues they are facing and recommend steps to remedy those issues. You are granted permission to send out a survey to all DCC students to collect data that will assist in identifying these students.

(a) Identify two categorical variables you might record about each student to determine whether the student is “at-risk”. For each variable, give two examples of data you might obtain.

(b) Identify two quantitative variables you might record about each student to determine whether the student is “at-risk”. For each variable, give two examples of data you might obtain.

12) Suppose that for the first exam in MAT185 (Precalculus) the instructor randomly chooses 20 students to take the exam while listening to instrumental music and 20 students to take the exam in silence. The instructor wants to determine whether the presence of music affects students’ exam scores.

(a) Identify (describe in words) the observational units and state the sample size (including the correct symbol) in this study.

(b) Identify (describe in words) one of the variables in this study and classify the variable as categorical or quantitative.

(c) Identify (describe in words) the other variable in this study and classify the variable as categorical or quantitative.

13) Recall that the key terms from Section 1.1 are observational units, variable (and variable type), population, sample (and sample size), data, parameter, and statistic. Describe in words what the key terms refer to in each of the following studies and which symbols (\( \hat{p}, \bar{x}, \mu, \pi, n \)) should be associated with those terms (if any). If possible, also give a value for each symbol.

(a) A group of National Fish and Wildlife researchers are collecting data about the bear population in Bearenstein National Park. To determine how environmental impacts are affecting the bears, the researchers collect data using the variables “sex” and “weight”. The researchers found that of the 29 bears surveyed, 16 were male and 13 were female, and the average weight of the bears was 692.5 lbs.

(b) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. Over a one-week period, 381 people across Dutchess County participated in the taste test and completed a short survey. PepsiCo found that the average number of soda beverages consumed per week is 5.7 and 205 of the 381 participants (205/381 = 0.54 = 54%) selected Pepsi as their preferred soda.

(c) You roll a twelve-sided die a total of 83 times and record the results of each roll. You then count the number of rolls where you obtained a ten or higher. Of the 83 rolls, a ten or higher was rolled a total of 26 times.

(d) A migraine is a particularly painful type of headache which people sometimes choose to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a study where 89 New York residents who were experiencing a migraine headache were either given acupuncture treatment designed to treat migraines, or they were given a “false” treatment (using acupuncture but in areas not intended to treat migraines). In a follow-up appointment the next day participants were asked whether they were now “pain-free”. Of the 89 participants, 43 received acupuncture intended to relieve migraines, and 46 did not. Of the 43 people who received the migraine treatment, 16 indicated they were pain-free after 24 hours. Of the 46 people who did not receive migraine treatment, 7 indicated they were pain-free after 24 hours.
(e) In 1948 the Chicago Daily Tribune erroneously published a headline due to early polling results that indicated Thomas E. Dewey had defeated Harry S. Truman in that year’s presidential election. The Gallup’s September 24 report indicated that they had obtained a random sample of 3,250 voters, and 46.5% of sampled voters indicated they would be voting for Dewey, while only 38% planned to vote for Truman. In truth, Truman won 49.6% of the popular vote (24,179,347 out of the 47,346,569 votes).

(f) We want to know the average (mean) amount of money that all first-year college students spend at the college bookstore on school supplies. We randomly survey 100 first-year students at Dutchess Community College. Three of those students spent $150, $200, and $225, respectively. We found that our sample had an average of $218.74 spent on school supplies.

(g) As part of a study designed to test the safety of automobiles, researchers at the National Transportation Safety Board (NTSB) collected and reviewed data about the effects of automobile crashes on test dummies. Following are the criteria they used:

<table>
<thead>
<tr>
<th>Speed at which cars crashed</th>
<th>Location of dummy</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 miles/hour</td>
<td>Front seat</td>
</tr>
</tbody>
</table>

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. The NTSB wants to know the proportion of all crashes in which the dummy in the driver’s seat would have had head injuries if it had been an actual driver. They start with a simple random sample of 75 crashes and find that in 15 of them the driver would have had head injuries.

14) According to their maker, different colors of Skittles candies are produced in equal quantities. However, they are packaged by weight and size, not color distribution. If I were to sample a pack of Skittles and record the proportion of each color, which of the following set of proportions is the most plausible (assuming different colors of Skittles candies are produced in equal quantities)? Explain your choice.

(a) Purple: 20%  Orange: 20%  Yellow: 20%  Red: 20%  Green: 20%
(b) Purple: 23%  Orange: 21%  Yellow: 21%  Red: 18%  Green: 17%
(c) Purple: 8%   Orange: 30%  Yellow: 27%  Red: 21%  Green: 4%
(d) Purple: 28%  Orange: 23%  Yellow: 24%  Red: 0%   Green: 25%

15) According to their maker, different colors of Skittles candies are produced in equal quantities. However, they are packaged by weight and size, not color distribution. Which of the following scenarios is more likely, or are they equally likely? Explain.

Scenario 1: You open a “fun size” bag of Skittles (approximately 16 Skittles) and find 75% of them are red.
Scenario 2: You open a “party size” bag of Skittles (approximately 1200 Skittles) and find that 75% of them are red.

16) A bucket of tickets contains 250 tickets of different colors. There are 75 blue tickets, 15 white tickets, and 160 green tickets. The tickets are thoroughly mixed in the bucket.

(a) What proportion of the tickets are blue?
(b) I am told that the probability of drawing a green ticket is 0.64. Explain what this means.
(c) On my first draw, I obtain a white ticket. What was the probability of this occurring?
Section 1.2 | Introduction to the Distribution of Sample Statistics

Key Terms

Simulation
Artificially re-creating a random process under a specific set of conditions. A larger number of repetitions gives us a better sense of the pattern we can expect among simulated sample statistics.

Distribution of Sample Statistics
A distribution of sample statistics that is created from a real or computerized process based on a set of assumptions or known information about the population parameter. We create a chance model by making an assumption about the true value of the population parameter or by actually knowing the true value of the population parameter, and then the distribution of sample statistics shows us the kinds of sample statistics we might expect to see if that assumption were true. We could create the distribution of sample statistics by hand (by flipping a fair coin, or spinning a fair spinner, etc.), or we could use a computer simulator like those found at https://www.rossmanchance.com/applets/index2021.html.

Mean of the Distribution of Sample Statistics
The value at the center of the distribution of sample statistics. It is the value that we are assuming to be the true value of the population parameter (or that we know to be the true value of the population parameter).

Normal Distribution
The normal distribution is a distribution that is “bell shaped” and often referred to as the “bell curve”. It is symmetric about the center value. The mean, median, and mode of the data are all located at the center of the distribution. 50% of the data is higher than the mean, and 50% is lower than the mean.

Standard Error (standard deviation of the distribution of sample statistics)
The average distance that the sample statistics are from the mean of the distribution. A big standard error means that there is a lot of spread in the distribution of sample statistics on average, and a small standard error means that the sample statistics we would expect to see would tend to stay closer to the mean on average.
Introduction to Simulation and the Distribution of Sample Statistics

In Section 1.1 we introduced the terms parameter and statistic. Recall that we generally do not know the value of the population parameter, but we do know the value of the statistic because we observed and calculated the value of the statistic from the sample data. We will now introduce the distribution of sample statistics, together with the concept of simulation, in order to begin to understand how an observed sample statistic can be used to help us gain a better understanding of the true value of an unknown population parameter. We will see that the concept of probability will play an important part in our reasoning.

**Simulation** is a process of artificially re-creating a random process under a specific set of conditions. We can use simulation to determine whether an assumed value of a population parameter is plausible/believable. **A distribution of sample statistics** is a dot plot of sample statistics created from a real or computerized process based on a set of assumptions. The following two collaborative exercises will give you the opportunity to participate in creating a simulated distribution of sample statistics. The first collaborative exercise explores simulation with quantitative data, and the second collaborative exercise explores simulation with categorical data.
In this Collaborative Exercise, we will perform an activity to estimate the average (mean) number of hours that all students at your college worked at a job last week.

(a) Have each student in the class count off in order to be assigned a student number (student #1, student #2, etc.). Each student should write down their assigned student number on a piece of paper that will be easily viewable to other students in the class. If possible, each student should pin or in some way attach that student number to his/her chest.

(b) Each student in the class should now individually go to the website www.random.org and generate five random numbers where the minimum number is 1, and the maximum number is the largest student number from part (a) above. Each student should write down the five random numbers generated from the website. Now each student in the class has five random numbers from part (b), as well as an assigned student number from part (a). Each student’s five random numbers identify the students in the class that each student needs to find and speak with.

(c) Once each student in the class completes part (b), the class should be asked to stand and walk around the room trying to find the students in the class who correspond to the five random numbers that they generated. Each “seeker” should seek out the five students on their list of random numbers. Once the seeker finds one of those students, the seeker should ask the student who is labeled with that random number to report how many hours that he/she worked at a job last week (it’s OK if the answer is 0). The seeker should write down the number of hours reported, and cross that student’s random number off their own list. The seeker should then walk around the room again in order to find the next of the five students on the list of random numbers. Repeat this until each student in the class has located all five students on their list of random numbers and written down the number of hours that each of those five students worked at a job last week. At the end, every student in the class should have five numbers written down which represent the number of hours worked last week by five random students in the class. Note that each student is a seeker, seeking five students in the class. But each student is also a student who will be searched for by other “seekers” in the class in order to report work hours.

(d) Everyone should return to their seats, and then calculate the average (mean) work hours for last week from the five students they talked to.

(e) As a class, the teacher will help to create a dot plot on the board at the front of the room by having each student in the class report the average number of hours worked by the five students they talked to. The dot plot that has been created is called a distribution of sample statistics.

(f) Is the data we are working with in this problem quantitative, or is it categorical? Explain.

(g) Each dot in the dot plot represents the average hours worked by five students in the class. What is the proper symbol for each dot? Also state whether each dot represents a parameter, or whether each dot represents a statistic.

(h) Describe the contextual meaning of the parameter of interest. Also identify the proper symbol for the parameter of interest.

(i) The parameter of interest has an unknown value. Use the dot plot created to give a best estimate of the true value of the parameter of interest. Explain why that is the estimate you gave. Discuss as a class.

(j) As a class, discuss how much confidence you have in the estimate you gave for the value of the population parameter. Explain. Also discuss and explain ways that we could potentially get a better estimate of the population parameter.

(k) If someone told you that the average (mean) number of hours worked by all students at your school last week was actually 36.2 hours, then what would you say to that person? Do you believe that is plausible/believable? Explain.
In this exercise, the students in the class will work together to verify the proportion of times, in the long run, that a fair die rolls the number one. The dice used in this activity are fair dice.

(a) Each student should roll a die 12 times and count the number of times (out of 12 rolls) that the die rolls the number one.

(b) Explain why we know the data from part (a) is from a categorical variable even though the calculated value (the count) is a number. Identify (describe in words) the variable. Identify (describe in words) the observational units.

(c) Each student should now calculate the proportion of times that their die rolled the number one out of their 12 rolls. If necessary, round the proportion to the hundredths place. Determine whether this calculated value is a statistic, or whether it is a parameter. Also give the correct symbol for this calculated value.

(d) Together as a class, create a dot plot of the distribution of sample statistics using each student’s calculated proportion. Be sure to label the contextual meaning of the horizontal axis.

(e) Describe the contextual meaning of the parameter of interest. Also give the correct symbol for the parameter of interest as well as its value (in this case the value of the parameter is known).

(f) The parameter of interest has a known theoretical value. Does the dot plot created in part (d) confirm the true value of the parameter of interest? Explain. Discuss as a class.

(g) Does the dot plot created by the class confirm that the dice that we rolled in the class are fair? Explain.

(h) As a class, discuss whether any students seem to have had a die that is not fair. How many ones out of 12 rolls would be convincing evidence of an unfair die?
Features of the Distribution of Sample Statistics (Chance Model)

In the two collaborative exercises above, we explored the creation of a distribution of sample statistics. It is important to understand the features of these distributions of sample statistics since they will play an important role in our statistical thinking.

**Dots in the Distribution of Sample Statistics**

Each dot in the distribution of sample statistics represents a statistic that occurred in the simulation when the value of the parameter is equal to an assumed (or known) value.

For example, in the “hours worked” collaborative exercise, each dot in the distribution of sample statistics actually represented a sample mean, $\bar{x}$, for a set of five students in the class. Note that we did not assume a value for the parameter, but used the dot plot to estimate the value of the parameter.

Remember that each dot is really representing a simulated sample statistic $\bar{x}$.

As another example, in the “roll a die” collaborative exercise, each dot in the distribution represents the proportion of times, $\hat{p}$, that a student rolled the number one when rolling the die 12 times. In this case, assuming the dice were fair, the value of the parameter is known to be $\pi = \frac{1}{6} = 0.167$.

Remember that each dot is really representing a simulated sample statistic $\hat{p}$.

**Shape of the Distribution of Sample Statistics**

As we increase the number of repetitions, the distribution of sample statistics will resemble a symmetric bell shape more and more closely. For example, in the “hours worked” collaborative exercise, if we had more and more students collect samples of 5 students all over campus, then we would see that the distribution of average hours worked more and more closely resembled a bell-shaped curve that has numerous sample statistics near the middle and a smaller number of sample statistics out in the “tails” of the distribution. In real-world terms, numerous samples of students would have an average number of work hours near the middle of the distribution, and only a small number of samples of students would have an average number of work hours that was a lot higher than that, or a lot lower than that.
Mean of the Distribution of Sample Statistics
The mean of the distribution is the value in the middle of the interval of sample statistics along the horizontal axis. The mean of the distribution will be close to the true value of the population parameter.

For example, in the “hours worked” collaborative exercise, the dots would be centered around the average hours worked last week for the entire population of students. So, the dots would be centered around the population mean, $\mu$. In the samples we collect from the population, about half of the sample means would be higher than the population mean, $\mu$, and about half of the sample means would be lower than the population mean, $\mu$.

For example, in the “roll a die” collaborative exercise, the dots would be centered around the long-run proportion of times the die would roll the number one. If the dice are fair, then the true value of the population parameter would be $\pi = \frac{1}{6}$ and so we would expect to usually get statistics that are near $1/6$, and only rarely get statistics much higher than $1/6$ or much lower than $1/6$. We would also expect about half of the statistics to be higher than $1/6$ and about half of the statistics to be lower than $1/6$. 
**Normal Distribution**
The normal distribution is a distribution that is “bell shaped” and often referred to as the “bell curve.” It is symmetric about the middle value. The middle value is the mean, median, and mode of the data (these values are all the same in a normal distribution). 50% of the data values are higher than the mean, and 50% are lower than the mean.

![Bell shaped curve](image)

**Standard Error (standard deviation of the distribution of sample statistics)**
The standard error is the average distance that the sample statistics are from the mean of the distribution. A big standard error means that there is a lot of spread in the distribution of sample statistics on average, and a small standard error means that the sample statistics are closer to the mean on average. In order to estimate the standard error, it is best to compare the distribution to the generic normal distribution in order to approximate the locations in the distribution where we would find data that is 1 standard deviation and 2 standard deviations above or below the mean. The picture of the normal distribution below indicates the location of data values that are in the middle (at the mean), and data values that are 1 standard deviation above/below the mean, 2 standard deviations above/below the mean, and 3 standard deviations above/below the mean.

![Standard error diagram](image)
There are three candy bars on the market that are in direct competition: Yummies, Crunchies, and Tasties. Yummies got 40% of all these candy bar purchases during the previous year. Suppose, during the previous year, a sample of 104 of those candy bar purchases were examined and the proportion of purchases that were Yummies was calculated.

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in a later section). The vertical lines are actually a large number of dots, which are so close together that they look like vertical lines. This distribution shows a set of simulated statistics that we would expect to see if it were true that Yummies were purchased 40% of the time in the population of all purchases. The distribution shows sample statistics when the size of each sample is 104.

Example 1.2.1

There are three candy bars on the market that are in direct competition: Yummies, Crunchies, and Tasties. Yummies got 40% of all these candy bar purchases during the previous year. Suppose, during the previous year, a sample of 104 of those candy bar purchases were examined and the proportion of purchases that were Yummies was calculated.

(a) Identify the contextual meaning of each dot in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) Estimate the mean of the distribution. Explain the contextual meaning of the mean.
(d) Estimate the standard deviation of the distribution. What is another name for the standard deviation of the distribution?

Solution 1.2.1

(a) Each dot in the distribution represents the sample proportion of purchases, \( \hat{p} \), that were Yummies in a simulated sample of 104 purchases in this population where 40% of all purchase are Yummies. This is a simulated distribution of sample p-hats.
(b) The shape of the distribution is generally bell-shaped, or approximately normal.
(c) The mean of the distribution is approximately 0.4. In context, this is the value of the population parameter, \( \pi = 0.4 \), which represents the fact that 40% of all purchases were Yummies in the previous year.
(d) The standard deviation is approximately 0.05. We get this approximation by comparing the simulated distribution to the markers for \( \pm 1 \) and \( \pm 2 \) standard deviations in the general normal distribution. Another name for the standard deviation of the distribution of sample statistics is the standard error. The standard error is approximately 0.05.
Example 1.2.2

Suppose the average (mean) GPA for all students at a certain college is 2.74. Suppose we collect a random sample of 85 students at the school and record the average (mean) GPA of this group of 85 students.

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in a later section). The little squares are just like the dots in a dot plot (this applet produces squares instead of dots). This distribution shows a set of sample means (from samples of size 85) that we might expect to see when the population mean GPA is 2.74.

Solution 1.2.2

(a) Identify the contextual meaning of each square in the distribution. Remember that the squares are just like the dots in a dot plot.

(b) Describe the shape of the distribution. Do the sample statistics appear to be roughly normally distributed?

(c) Estimate the mean of the distribution. Explain the contextual meaning of the mean.

(d) Estimate the standard deviation of the distribution. What is another name for the standard deviation of the distribution?

Example 1.2.2

Suppose the average (mean) GPA for all students at a certain college is 2.74. Suppose we collect a random sample of 85 students at the school and record the average (mean) GPA of this group of 85 students.

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in a later section). The little squares are just like the dots in a dot plot (this applet produces squares instead of dots). This distribution shows a set of sample means (from samples of size 85) that we might expect to see when the population mean GPA is 2.74.

(a) Identify the contextual meaning of each square in the distribution. Remember that the squares are just like the dots in a dot plot.

(b) Describe the shape of the distribution. Do the sample statistics appear to be roughly normally distributed?

(c) Estimate the mean of the distribution. Explain the contextual meaning of the mean.

(d) Estimate the standard deviation of the distribution. What is another name for the standard deviation of the distribution?
Section 1.2 | Introduction to the Distribution of Sample Statistics

Exercises

1) **(Multiple Choice)** Suppose that Rob records the ages of a random sample of 1500 students at SUNY Albany, while Linda records the ages of a random sample of 1500 residents in Poughkeepsie. Whose data would you expect to have the larger standard deviation of ages?
   (a) Linda’s data would likely have a higher standard deviation than Rob’s data.
   (b) Rob’s data would likely have a higher standard deviation than Linda’s data.
   (c) They would likely have the same standard deviation for age because they have the same sample size.
   (d) There is not enough information to tell.

2) **(Multiple Choice)** Frank records the monthly rent for a random sample of 100 apartments in Manhattan, while Theresa does the same in Poughkeepsie. Whose data would you expect to have a higher mean of rents?
   (a) Frank’s data would likely have a higher mean rent amount than Theresa’s data.
   (b) Theresa’s data would likely have a higher mean rent amount than Frank’s data.
   (c) They would likely have the same mean rent amount because they have the same sample size.
   (d) There is not enough information to tell.

3) **(Multiple Choice)** Brian records the quantitative SAT score for a random sample of students attending Ivy League colleges, while Felicia does the same for a random sample of college students. Whose data would you expect to have the larger standard deviation of SAT scores?
   (a) Brian’s data would likely have a higher standard deviation than Felicia’s data.
   (b) Felicia’s data would likely have a higher standard deviation than Brian’s data.
   (c) They would likely have the same standard deviation because they have the same sample size.
   (d) There is not enough information to tell.

4) Suppose Ben records the noon temperature in New York City every day in the month of July, and Jerry records the noon temperature in New York City every day in an entire year.
   (a) Which one (Ben or Jerry) would you expect to have the smaller mean of temperatures, or would you expect the means to be very similar? Explain briefly.
   (b) Which one (Ben or Jerry) would you expect to have the smaller standard deviation of temperatures, or would you expect the standard deviations to be very similar? Explain briefly.

5) Suppose that Jimmy records the heights of all residents of Poughkeepsie, while Julie records the heights of all NBA basketball players. Who would you expect to have the larger standard deviation of heights? Explain.

6) Suppose that Allie records the ages of all fifth-grade students in the Hyde Park school district, while Evan records the ages of all Hyde Park residents. Who would you expect to have the larger standard deviation of ages? Explain.

7) The graph of the distribution of simulated sample statistics will be centered approximately at ____.
   (a) the observed statistic
   (b) the observed count
   (c) the population parameter
   (d) the number of repetitions performed
8) LeBron James, formerly of the Miami Heat, hit 765 of his 1354 field goal attempts in the 2012/2013 season for a shooting percentage of 56.5%. Could this have happened by chance? Which of the following is a correct description of the parameter of interest?
(a) LeBron’s average shooting percentage if we look at several sets of 1354 field goal attempts.
(b) LeBron’s shooting percentage for the 2012/2013 season.
(c) LeBron’s shooting percentage throughout his career.
(d) LeBron’s shooting percentage if he has a 50% chance of making any particular field goal.

9) I am arguing with a fellow coffee connoisseur about whether the average coffee drinker can tell the difference between a latte and a macchiato (Note: This is a big deal to serious espresso drinkers). Suppose that I sample 7 espresso drinks and correctly identify 6 of them as either a latte or a macchiato.
(a) Describe the parameter of interest in the context of this study. What symbol should be used to represent the parameter?
(b) What is the observed value of the statistic in this case? What symbol should be used to represent the statistic?

10) A study is being conducted to determine whether Nature-All (a new natural, organic, aluminum-free deodorant) can perform as well as the industry standard antiperspirant (containing aluminum) at preventing body odor. Forty people apply Nature-All to their bodies and then participate in a 1-hour high-intensity workout. A team of researchers then smells the participant and records at what distance (in feet from the participant) they can begin to smell the participant’s body odor.
(a) Describe the parameter of interest in the context of this study. What symbol should be used to represent the parameter?
(b) What is the observed value of the statistic in this case? What symbol should be used to represent the statistic?

11) The heights of the NFL 2020-2021 season wide receivers and quarterbacks were recorded. Dot plots of the data are below.

Note: Your answers and explanations to the following questions should be based only on what you see in the dot plots, not on conjectures about why or how the sport played might affect a person’s height.
(a) Without doing any calculations, which group’s data would you expect to have the larger mean height? Explain.
(b) Without doing any calculations, which group’s data would you expect to have the larger standard deviation? Explain.
12) A researcher asked students at Dutchess Community College whether they are full-time students (enrolled in 12 or more credit hours for the fall or spring semester) or part-time students (enrolled in 11 or fewer credits). The researcher also recorded how many media-streaming subscriptions (such as Netflix) are held by the student’s household. The resulting dot plots are shown below.

Note: Your answers and explanations should be based only on what you see in the dot plots, not on conjectures about why or how enrollment status might affect the number of media-streaming subscriptions.

(a) Without doing any calculations, which group’s data would you expect to have the larger mean of media-streaming subscriptions? Explain.
(b) Without doing any calculations, which group’s data would you expect to have the larger standard deviation of media-streaming subscriptions? Explain.

13) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that 2 out of every 3 people prefer their brand of cola. A random sample of 80 people is given the taste test, and 45 of them (56%) choose Pepsi as their preferred cola.

To conduct a statistical analysis, a distribution of 300 simulated sample statistics was generated using a computer applet as seen below (we will learn how to do this in a later section). This distribution shows the statistics that we might expect to see when randomly sampling 80 people if it were true that 2 out of 3 people prefer Pepsi brand cola.

(a) Identify the contextual meaning of each dot in the distribution.
(b) Describe the shape of the distribution. Does it appear to be roughly normally distributed?
(c) Estimate the mean of the distribution. Explain the contextual meaning of the mean.
(d) Estimate the standard deviation of the distribution. What is another name for this value?
Kelley Blue Book is a vehicle valuation and automotive research company that reports market value prices for new and used automobiles. These values are often used by auto dealerships to determine the selling price of their vehicles. The average Kelley Blue Book value of a used 2008-2012 sedan is $5,560. I would like to determine if I can ask for a different selling price in my region for my sedan, so I conduct a random sample of local ads for used 2008-2012 sedans and record the asking price for 28 of them, obtaining an average asking price of $5,985.

To conduct a statistical analysis, a distribution of simulated sample statistics was generated using a computer applet as seen below (we will learn how to do this in a later section). The little squares are just like the dots in a dot plot (this applet produces squares instead of dots). This distribution shows a set of statistics that we might expect to see if the population mean is 5560.

(a) Identify the contextual meaning of each square in the distribution.
(b) Describe the shape of the distribution. Do the simulated statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of that mean.
(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
In 1948 the Chicago Daily Tribune erroneously published a headline due to early polling results that indicated Thomas E. Dewey had defeated Harry S. Truman in that year’s presidential election. In truth, Truman won 49.6% of the popular vote (24,179,347 out of 47,346,569 votes) to Dewey’s 45.1% (21,991,292 out of 47,346,569 votes) with the remaining percentage going to third party candidates.

The Gallup’s September 24 report indicated that they had obtained a random sample of 3,250 voters, and 46.5% of sampled voters indicated they would be voting for Dewey, while only 38% planned to vote for Truman.

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in a later section). This distribution shows a set of statistics that we might expect to see if we surveyed 3,250 people from a population where 45.1% of the voting population will vote for Dewey.

(a) Identify the contextual meaning of each dot in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of the mean.
(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
(e) Given that we now know Dewey received 45.1% of the popular vote, was Gallup’s statistic (46.5%) reasonable according to the distribution of simulated statistics? Explain.

Now consider the same Gallup poll from the perspective of Harry S. Truman’s campaign team (the poll indicated he would obtain 38% of the vote). Recall that in reality, Truman won 49.6% of the popular vote. The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in a later section). This distribution shows a set of statistics that we might expect to see if we surveyed 3,250 people from a population where 49.6% of the voting population will vote for Truman.

(f) Identify the contextual meaning of each dot in the distribution.
(g) Describe the shape of the distribution. Do the sample statistics appear to be roughly normally distributed?
(h) What is the mean of the distribution? Explain the contextual meaning of the mean.
(i) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
(j) Given that we now know Truman received 49.6% of the popular vote, was Gallup’s statistic (38%) reasonable according to the distribution of simulated statistics? Explain.
**Section 1.3 | Understanding when Statistics are Unusual**

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
</table>
| • Fluently use the following vocabulary and symbols:  
  O 5% Rule  
  O Empirical Rule  
  O 2 Standard Deviations Rule  
• Use the Empirical Rule to answer questions about the percentage of statistics we expect to find at various distances from the population parameter.  
• Consider a distribution of sample statistics and identify sample statistics that are “usual and expected” vs. “unusually high or unusually low” using the 2 Standard Deviations Rule and given information about the true value of the population parameter. |

**Key Terms**

**5% Rule**  
Events that have a probability that is less than 5% are considered “unusual”.

**Empirical Rule (68-95-99.7 Rule)**  
If a distribution of sample statistics is normally distributed, then  
• About 68% of the sample statistics lie from 1 standard deviation below the mean of the distribution up to 1 standard deviation above the mean of the distribution.  
• About 95% of the sample statistics lie from 2 standard deviations below the mean of the distribution up to 2 standard deviations above the mean of the distribution.  
• About 99.7% of the sample statistics lie from 3 standard deviations below the mean of the distribution up to 3 standard deviations above the mean of the distribution.

**2 Standard Deviations Rule**  
If a distribution of sample statistics is normally distributed, then  
• It is “unusual” for a statistic to be more than 2 standard deviations above the mean of the distribution: such statistics would generally be considered “unusually high”.  
• It is unusual for a statistic to be lower than 2 standard deviations below the mean of the distribution: such statistics would generally be considered “unusually low”.  
• Only about 5% of the statistics are more than 2 standard deviations away from the mean of the distribution: such statistics are generally considered “unusual”.

**5% Rule**  
Events that have a probability that is less than 5% are considered unusual, and events that have a probability that is more than 5% are considered usual.

Please note that we are only defining 5% to be the “cut-off” for usual/expected vs. unusual at this point in the book in order to allow us to have some common language to use until we learn some additional statistical vocabulary, notation, and concepts. Later in the book we will use different, more commonly used standards for concluding whether a statistic is usual/expected vs. unusual/unexpected. For example, it is common to use a p-value and significance level in order to make such conclusions.
The Empirical Rule and the 2 Standard Deviations Rule

If we assume or know the true value of a population parameter, then the distribution of simulated sample statistics (also called a chance model) allows us to see the sorts of sample statistics that we might expect from samples of a particular size, together with the frequency of those sample statistics. If a sample statistic occurs infrequently or not at all in a chance model, then we would consider it an unusual/unexpected sample statistic. If a sample statistic occurs frequently in a chance model, the we would consider it a usual/expected sample statistic.

Unfortunately, the terms “frequently” and “infrequently” are generic and broad. Such non-specific language would lead different people to come to different conclusions about which statistics are usual vs. unusual. The Empirical Rule and the 2 Standard Deviation Rule will allow us to create common conclusions about which statistics are usual vs. unusual.

**Empirical Rule**

If a distribution of sample statistics is normally distributed, then

- About 68% of the sample statistics lie from 1 standard deviation below the mean of the distribution up to 1 standard deviation above the mean of the distribution (left picture below).
- About 95% of the sample statistics lie from 2 standard deviations below the mean of the distribution up to 2 standard deviations above the mean of the distribution (middle picture below).
- About 99.7% of the sample statistics lie from 3 standard deviations below the mean of the distribution up to 3 standard deviations above the mean of the distribution (right picture below).
2 Standard Deviations Rule
If a distribution of sample statistics is normally distributed, then

- It is unusual for a statistic to be more than 2 standard deviations above the mean of the distribution. Such statistics would generally be considered unusually high.
- It is unusual for a statistic to be lower than 2 standard deviations below the mean of the distribution. Such statistics would generally be considered unusually low.
- Only about 5% of the statistics are more than 2 standard deviations away from the mean of the distribution. Such statistics are generally considered unusual.

This 2 Standard Deviations Rule follows from the Empirical Rule. Notice that the Empirical Rule tells us that about 95% of the sample statistics lie from 2 standard deviations below the mean of the distribution up to 2 standard deviations above the mean of the distribution. It follows, then, that about 2.5% of the statistics would be higher than 2 standard deviations above the mean, and about 2.5% of the statistics would be lower than 2 standard deviations below the mean of the distribution.
There are three candy bars on the market that are in direct competition: Yummies, Crunchies, and Tasties. Yummies got 40% of all these candy bar purchases during the previous year ($\pi = 0.4$). Suppose, during the previous year, a sample of 104 of those candy bar purchases were examined and the proportion of purchases that were Yummies was calculated.

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in the next section). The vertical lines are actually a large number of dots, which are so close together that they look like vertical lines. This distribution shows the simulated statistics that we would expect to see if it were true that Yummies were purchased 40% of the time in the population of all purchases. The distribution shows sample statistics when the size of each sample is 104. Notice that the statistics appear to be normally distributed, and that the top left-hand corner displays the mean and standard deviation (SD) of this simulated distribution.

(a) Use the Empirical Rule to complete the sentence:
About ________% of the samples of 104 purchases had statistics where more than 49.8% of the 104 people purchased Yummies.

(b) Use the 2 Standard Deviations Rule to complete the sentence:
It would be unusual to have a sample of 104 purchases where more than ________% of the purchases were Yummies.

(c) Use the 2 Standard Deviations Rule to complete the sentence:
It would be unusual to have a sample of 104 purchases where fewer than ________% of the purchases were Yummies.

(d) Use the 2 Standard Deviations Rule to complete the sentence:
When considering samples of 104 purchases, the usual/expected statistics come from samples where between ________% and ________% of the purchases are Yummies.

Solution 1.3.1

(a) The mean is shown in the picture as 0.4 = 40%, which matches the given parameter value, $\pi = 0.4$. The standard deviation is shown in the picture as 0.049 = 4.9%. 1 standard deviation above the mean is 40% + 4.9% = 44.9%. 2 standard deviations above the mean is 40% + 4.9% + 4.9% = 49.8%. According to the Empirical Rule, only about 5% of the statistics are more than 2 standard deviations away from the mean: 2.5% on the “top end” and 2.5% on the “low end”. This means that

About ________% of the samples of 104 purchases had statistics where more than 49.8% of the 104 people purchased Yummies.

(continued on next page)
Solution 1.3.1 continued

(b) According to the 2 Standard Deviations Rule, “unusual” means that the sample statistic is more than 2 standard deviations above or below the mean of the distribution. This means that the statistic would need to be more than $0.4 + 0.049 + 0.049 = 0.498$, or less than $0.4 - 0.049 - 0.049 = 0.302$.

It would be unusual to have a sample of 104 purchases where more than $49.8\%$ of the purchases were Yummies.

(c) According to the 2 Standard Deviations Rule, “unusual” means that the sample statistic is more than 2 standard deviations above or below the mean of the distribution. This means that the statistic would need to be more than $0.4 + 0.049 + 0.049 = 0.498$, or less than $0.4 - 0.049 - 0.049 = 0.302$.

It would be unusual to have a sample of 104 purchases where fewer than $30.2\%$ of the purchases were Yummies.

(d) As we worked out above, the “cut-offs” for usual vs. unusual, according the 2 Standard Deviations Rule, are the statistics 0.498 on the upper end, and 0.302 on the lower end.

When considering samples of 104 purchases, the usual/expected statistics come from samples where between $30.2\%$ and $49.8\%$ of the purchases are Yummies.

The distribution of simulated sample p-hats can be correlated to the generic distribution of sample p-hats, which is shown below. Unusual sample statistics are more than 2 standard deviations from the center (mean) of the distribution, or in the “tails” of the distribution.
Example 1.3.2

Suppose the average (mean) GPA for all students at a certain college is 2.74 (μ = 2.74). Suppose we collect a sample of 85 students at the school and record the average (mean) GPA of each group of 85 students.

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in the next section). The little squares are just like the dots in a dot plot (this applet produces squares instead of dots). This distribution shows a set of sample means (from samples of size 85) that we might expect to see when the population mean GPA is 2.74. Notice that the statistics are approximately normally distributed, and that the top left-hand corner displays the mean and standard deviation (SD) of this simulated distribution.

(a) Use the Empirical Rule to complete the sentence:
   About ________% of the samples of 85 students had statistics where the average GPA was less than 2.718.

(b) Use the 2 Standard Deviations Rule to complete the sentence:
   It would be unusual to have a sample of 85 students in which the average GPA is more than ________.

(c) Use the 2 Standard Deviations Rule to complete the sentence:
   It would be unusual to have a sample of 85 students in which the average GPA is less than ________.

(d) Use the 2 Standard Deviations Rule to complete the sentence:
   When considering samples of 85 students, the usual/expected average GPAs are those that are between ________ and ________.

Solution 1.3.2

(a) The mean is shown in the picture as 2.737, which coincides with the given population mean GPA of μ = 2.74. The standard deviation is shown in the picture as 0.019. 1 standard deviation below the mean is 2.737 − 0.019 = 2.718. According to the Empirical Rule, about 32% of the statistics are more than 1 standard deviation away from the mean: 16% on the “top end” and 16% on the “low end”. This means that

   About ____16____% of the samples of 85 students had statistics where the average GPA was less than 2.718.

(continued on next page)
Solution 1.3.2 continued

(b) According to the 2 Standard Deviations Rule, “unusual” means that the sample statistic is more than 2 standard deviations above or below the mean of the distribution. This means that the statistic would need to be more than \(2.737 + 0.019 + 0.019 = 2.775\), or less than \(2.737 - 0.019 - 0.019 = 2.699\).

It would be unusual to have a sample of 85 students in which the average GPA is more than \(2.775\).

(c) According to the 2 Standard Deviations Rule, “unusual” means that the sample statistic is more than 2 standard deviations above or below the mean of the distribution. This means that the statistic would need to be more than \(2.737 + 0.019 + 0.019 = 2.775\), or less than \(2.737 - 0.019 - 0.019 = 2.699\).

It would be unusual to have a sample of 85 students in which the average GPA is less than \(2.699\).

(d) As we worked out above, the “cut-offs” for usual vs. unusual, according the 2 Standard Deviations Rule, are the statistics \(2.775\) on the upper end, and \(2.699\) on the lower end.

When considering samples of 85 students, the usual/expected average GPAs are those that are between \(2.699\) and \(2.775\).

The distribution of simulated sample x-bars can be correlated to the generic distribution of sample x-bars, which is shown below. Unusual sample statistics are more than 2 standard deviations from the center (mean) of the distribution, or in the “tails” of the distribution.
Section 1.3 | Understanding when Statistics are Unusual

Exercises

For questions 1 through 3, use the Empirical Rule (68-95-99.7 Rule) to determine the appropriate value. It can be helpful to display a graph with the values from the Empirical Rule labeled to visualize the important areas.

1) If a distribution of sample statistics is normally distributed, what percentage of sample statistics will be more than 2 standard deviations away from the mean?

(a) 99.7%  (c) 68%  (e) 16%  (g) 5%  (i) 0.3%
(b) 95%  (d) 32%  (f) 8%  (h) 2.5%  (j) 0.15%

2) If a distribution of sample statistics is normally distributed, what percentage of sample statistics will be more than 1 standard deviation above the mean?

(a) 99.7%  (c) 68%  (e) 16%  (g) 5%  (i) 0.3%
(b) 95%  (d) 32%  (f) 8%  (h) 2.5%  (j) 0.15%

3) If a distribution of sample statistics is normally distributed, what percentage of sample statistics will be more than 3 standard deviations below the mean?

(a) 99.7%  (c) 68%  (e) 16%  (g) 5%  (i) 0.3%
(b) 95%  (d) 32%  (f) 8%  (h) 2.5%  (j) 0.15%

4) If a distribution of sample statistics is normally distributed, what percentage of sample statistics will be less than 2 standard deviations away from the mean?

(a) 99.7%  (c) 68%  (e) 16%  (g) 5%  (i) 0.3%
(b) 95%  (d) 32%  (f) 8%  (h) 2.5%  (j) 0.15%

5) If a distribution of sample statistics is normally distributed, what percentage of sample statistics will be within 3 standard deviations of the mean?

(a) 99.7%  (c) 68%  (e) 16%  (g) 5%  (i) 0.3%
(b) 95%  (d) 32%  (f) 8%  (h) 2.5%  (j) 0.15%

6) A distribution of sample statistics is normally distributed with a mean of 32 and a standard deviation of 1.9. Use the Empirical Rule to complete the following sentences:

(a) ________ % of the sample statistics are greater than 37.7.
(b) It would be unusual to obtain a sample statistic greater than ____________.
(c) It would be unusual to obtain a sample statistic less than ____________.
(d) 68% of the sample statistics are between ____________ and ____________.
7) A distribution of sample statistics is normally distributed with a mean of 0.333 and a standard deviation of 0.085. Use the Empirical Rule to complete the following sentences:
(a) _________% of the sample statistics are less than 0.248.
(b) It would be unusual to obtain a sample statistic greater than ______________.
(c) It would be unusual to obtain a sample statistic less than ______________.
(d) 99.7% of the sample statistics are between _____________ and ____________.

8) Suppose that Erick and Nina both collect data to investigate whether people tend to call “heads” more often than “tails” when they are asked to call the result of a coin flip. Erick and Nina collect their sample data independently, then compare their sample proportions and find that they obtained the same sample proportion of 0.624 (or 62.4%).

(a) After observing this, Erick claims that their two studies provide the same strength of evidence for the statement “people tend to call heads more than tails”. Explain why Erick is not necessarily correct.

(b) Can you make any conclusion at this point regarding people’s tendency to pick heads more often than tails? What is needed to make a more definitive statement?

(c) The distribution of simulated sample statistics generated by Erick is shown directly to the right. Is Erick’s statistic usual or unusual? What conclusion should Erick draw based on his sample proportion? Be sure to justify your answer.

(d) The distribution of simulated sample statistics generated by Nina is shown directly to the right. Is Nina’s statistic usual or unusual? What conclusion should Nina draw based on her sample proportion? Be sure to justify your answer.
The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that 2 out of every 3 people prefer their brand of cola. A random sample of 80 people is taken, and 45 of them (56%) choose Pepsi as their preferred cola.

To conduct a statistical analysis, a distribution of simulated sample statistics was generated below using a computer applet (we will learn how to do this in the next section). The vertical lines are a large number of dots, which are so close together that they look like vertical lines. The distribution shows the simulated statistics that we would expect to see when randomly sampling 80 people if it were true that 2 out of 3 people prefer Pepsi brand cola. Notice that the top left-hand corner displays the mean and standard deviation (SD) of this distribution of simulated sample statistics.

![Distribution of simulated sample statistics](image)

Proportion of Participants with a Preference for Pepsi

(a) Identify the contextual meaning of each dot in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of this mean.
(d) What is the standard error? What is another name for the standard error?
(e) Use the 2 Standard Deviations Rule to determine the cutoffs for “unusual” statistics.
(f) Use the results from part (e) to complete the following statements:

If 2 out of every 3 people prefer Pepsi over Coca-Cola, it would be unusual to sample 80 people and obtain a sample proportion that is less than __________.

If 2 out of every 3 people prefer Pepsi over Coca-Cola, it would be unusual to sample 80 people and obtain a sample proportion that is greater than __________.

If 2 out of every 3 people prefer Pepsi over Coca-Cola, then the usual/expected proportion who prefer Pepsi over Coca-Cola in a sample of 80 people is between __________ and __________.
10) A study was conducted by the National Endowment for the Arts to describe music preferences in the Hudson Valley. The 2010 report described relationships between key demographic characteristics and music preferences. The study found that 24% of adults preferred the genre of Rock/Heavy Metal at that time. Recently, a random sample of 715 Hudson Valley adults was surveyed, with 150 indicating they prefer Rock/Heavy Metal. You want to determine whether the proportion of Hudson Valley residents that prefer Rock/Heavy Metal has changed since the 2010 report.

To conduct a statistical analysis, a distribution of 1000 simulated sample statistics was generated below using a computer applet (we will learn how to do this in the next section). The distribution shows the simulated statistics that we would expect to see when surveying 715 people, if it were true that 24% of Hudson Valley residents prefer Rock/Heavy Metal. Notice that the top left-hand corner displays the mean and standard deviation (SD) of this distribution of simulated sample statistics.

(a) Identify the contextual meaning of each dot in the distribution.

(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?

(c) What is the mean of the distribution? Explain the contextual meaning of this mean.

(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?

(e) Use the 2 Standard Deviations Rule to calculate the cutoffs for unusual statistics.

(f) Use the results from part (e) to complete the following statements:

If 24% of the population in the Hudson Valley prefers Rock/Heavy Metal music, it would be unusual to sample 715 Hudson Valley residents and find that fewer than ________% of those surveyed prefer Rock/Heavy Metal music.

If 24% of the population in the Hudson Valley prefers Rock/Heavy Metal music, it would be unusual to sample 715 Hudson Valley residents and find that more than ________% of those surveyed prefer Rock/Heavy Metal music.

If 24% of the population in the Hudson Valley prefers Rock/Heavy Metal music, then the usual/expected percentage of residents who prefer Rock/Heavy Metal from a random sample of 715 Hudson Valley residents is between ________ and ________.
11) *PriceCars4You.com* is a website used to buy and sell automobiles. Insurance companies often use these websites with an algorithm to estimate the value of your car when you file a claim. Recently, I was rear-ended, and our insurance company (Progressive) estimated the value of our 2010 Hyundai Sonata to be $3200.00. They claimed this value is based on the average selling price of similar cars in the area. I would like to research the average value of 2010 Hyundai Sonatas in my area to determine whether I am being offered a fair price, or if I should argue for a higher value. Progressive stated that their sample used only 4 cars, so I conduct a random sample of local ads for used 2010 Hyundai Sonatas with comparable mileage and record the asking price for 4 of them, obtaining an average asking price of $4,150.00.

The distribution of simulated sample statistics to the right was generated using a computer applet to model the situation (we will learn how to do this in the next section). The little squares are just like the dots in a dot plot (this applet produces squares instead of dots). The distribution displays statistics we would expect to see when sampling 4 cars if the average asking price for all 2010 Hyundai Sonatas with comparable mileage in my area is $3200.00.

(a) Identify the contextual meaning of each square in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of this mean.
(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
(e) Use the 2 Standard Deviations Rule to calculate the cutoffs for unusual statistics.
(f) Use the results from part (e) to complete the following statements:

If the average asking price in my region for a 2010 Hyundai Sonata is $3,200.00, it would be unusual to obtain a random sample of 4 vehicle ads where the average vehicle asking price is less than ___________.

If the average asking price in my region for a 2010 Hyundai Sonata is $3,200.00, it would be unusual to obtain a random sample of 4 vehicle ads where the average vehicle asking price is greater than ___________.

If the average asking price in my region for a 2010 Hyundai Sonata is $3,200.00, an expected/usual average asking price from a sample of 4 advertisements selling 2010 Hyundai Sonatas would be between __________ and ___________.

(g) Should I argue for a better price on my totaled 2010 Hyundai Sonata? Explain your reasoning using your work in parts (a) through (f).
In 1948 the Chicago Daily Tribune erroneously published a headline due to early polling results that indicated Thomas E. Dewey had defeated Harry S. Truman in that year’s presidential election. In truth, Truman won 49.6% of the popular vote (24,179,347 out of 47,346,569 votes) to Dewey’s 45.1% (21,991,292 out of 47,346,569 votes) with the remaining percentage going to third party candidates.

The Gallup’s September 24 report indicated that they had obtained a random sample of 3,250 voters, and 46.5% of sampled voters indicated they would be voting for Dewey while only 38% planned to vote for Truman.

The distribution of simulated sample statistics to the right was generated using a computer applet to model the situation (we will learn how to do this in the next section). The distribution shows a set of simulated statistics that we might expect to see if Dewey got 45.1% of the popular vote, and we then collected samples and recorded the percentage of people who voted for Dewey in each sample.

(a) Identify the contextual meaning of each dot in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of this mean.
(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
(e) Use the 2 Standard Deviations Rule to calculate the cutoffs for unusual statistics.
(f) Use the results from part (e) to complete the following statements:

Knowing now that 45.1% of people voted for Dewey, it would be unusual to obtain a sample of 3,250 people and find that fewer than % planned vote for Dewey.

Knowing now that 45.1% of people voted for Dewey, it would be unusual to obtain a sample of 3,250 people and find that more than % planned to vote for Dewey.

Knowing now that 45.1% of people voted for Dewey, a usual/expected percentage of people planning to vote for Dewey in a sample of 3,250 people would be between and .

(#12 continues on the next page)
Now consider the same Gallup poll from the perspective of Harry S. Truman’s campaign team. Recall that Truman won 49.6% of the popular vote, while the poll indicated he would obtain 38% of the vote. The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in the next section). The distribution shows a set of statistics that we might expect to see if we surveyed 3,250 people from a population where 49.6% of the voting population will vote for Truman.

(g) Identify the contextual meaning of each dot in the distribution.
(h) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(i) What is the mean of the distribution? Explain the contextual meaning of this mean.
(j) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution of simulated sample statistics?
(k) Use the 2 Standard Deviations Rule to calculate the cutoffs for unusual statistics.
(l) Use the results from part (e) to complete the following statements:

- Knowing now that 49.6% of people voted for Truman, it would be unusual to obtain a sample of 3,250 people and find that fewer than __________ % planned to vote for Truman.
- Knowing now that 49.6% of people voted for Truman, it would be unusual to obtain a sample of 3,250 people and find that more than __________ % planned to vote for Truman.
- Knowing now that 49.6% of people voted for Truman, a usual/expected percentage of people planning to vote for Truman in a sample of 3,250 people would be between _________ and __________.

(m) **Applying our analysis**: Given that the Gallup poll was published only a week prior to the election (and therefore conducted shortly before that), what is a plausible explanation for the discrepancy between the polls and the election results?
13) The United States Health and Human Services Poverty Guidelines for 2017 are displayed below:

<table>
<thead>
<tr>
<th>Family/Household Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty Threshold</td>
<td>$12,060</td>
<td>$16,240</td>
<td>$20,420</td>
<td>$24,600</td>
<td>$28,780</td>
<td>$32,960</td>
<td>$37,140</td>
</tr>
</tbody>
</table>

In March 2017 a PBS report\(^1\) stated that approximately 8.3% of citizens live below the poverty line in the United States of America. A study is conducted to determine how New York State (NYS) residents compare to these statistics. In a random sample of 10,000 NYS residents, it is determined that 1,361 (13.61%) were living below the poverty line in 2019. We want to determine the probability of obtaining a sample proportion of 0.1361 from a sample of 10,000 NYS residents if the true proportion of all NYS residents living below the poverty line is 0.083.

To conduct a statistical analysis, a distribution of simulated sample statistics was generated below using a computer applet (we will learn how to do this in the next section). The distribution shows the statistics that we would expect to see when surveying 10,000 NYS residents if 8.3% of all NYS residents live below the poverty line. Notice that the top left-hand corner displays the mean and standard deviation of this distribution of simulated sample statistics, and the horizontal axis is showing the number (or count), not a proportion.

(a) Identify the contextual meaning of each dot in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of this mean.
(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
(e) Use the 2 Standard Deviations Rule to calculate the cutoffs for unusual statistics.
(f) Use the results from part (e) to complete the following statements:

If 8.3% of the population in New York State lives below the poverty line, it would be unusual to obtain a random sample of 10,000 people where fewer than ________________ people live below the poverty line.

If 8.3% of the population in New York State lives below the poverty line, it would be unusual to obtain a random sample of 10,000 people where more than ________________ people live below the poverty line.

If 8.3% of the population in New York State lives below the poverty line, then the usual/expected number of residents who live below the poverty line in a random sample of 10,000 NYS residents would be between _________ and ___________.

\(^1\) https://www.pbs.org/newshour/nation/six-charts-illustrate-divide-rural-urban-america
14) The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years. After learning this you decide to collect data from a random sample of New York City (NYC) obituaries from 2018-2020 to compare the life expectancy of New York City residents to the life expectancy of all U.S. citizens.

After surveying 75 obituaries, you find a mean life expectancy of 80.7 years. Is this a reasonable sample mean to obtain if the average life expectancy of all NYC residents is 78.7 years?

The distribution of simulated sample statistics below was generated using a computer applet to model the situation (we will learn how to do this in the next section). The little squares are just like the dots in a dot plot (this applet produces squares instead of dots). The distribution shows a set of statistics that we would expect to see from a random sample of 75 obituaries if the true average life expectancy for all NYC residents is 78.7 years.

(a) Identify the contextual meaning of each square in the distribution.
(b) Describe the shape of the distribution. Do the simulated sample statistics appear to be roughly normally distributed?
(c) What is the mean of the distribution? Explain the contextual meaning of this mean.
(d) What is the standard deviation of the distribution? What is another name for the standard deviation of the distribution?
(e) Use the 2 Standard Deviations Rule to calculate the cutoffs for unusual statistics.
(f) Use the results from part (e) to complete the following statements:

If the average life expectancy of all NYC residents is 78.7 years, it would be unusual to obtain a sample of 75 obituaries of NYC residents where the average age at the time of death is less than _________ years.

If the average life expectancy of all NYC residents is 78.7 years, it would be unusual to obtain a sample of 75 obituaries of NYC residents where the average age at the time of death is greater than _________ years.

If the average life expectancy of all NYC residents is 78.7 years, then the usual/expected average age at the time of death from a sample of 75 NYC obituaries would be between _________ and _________ years.

(g) Given the statements from part (f), is it reasonable to believe that New York City residents have, on average, an unusually long life expectancy compared to the rest of the country?

1 https://www.cdc.gov/nchs/products/databriefs/db355.htm
Section 1.4 | Introduction to Computer Simulation

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fluently use the following vocabulary and symbols:</td>
</tr>
<tr>
<td>O Computer simulation</td>
</tr>
<tr>
<td>• For problems involving a proportion, use a computer simulation to create a distribution of the statistics we would expect to see based on assumed or known information about the population parameter.</td>
</tr>
<tr>
<td>• Interpret the contextual meaning of the features of a distribution of simulated sample statistics for a mean, or a distribution of simulated sample statistics for a proportion.</td>
</tr>
<tr>
<td>• Use the computer applet, or a given pictured distribution from an applet, to identify the percentage of statistics that are at least as extreme as a given value. Use the percentage to identify if the observed statistic is “usual” vs. “unusual.”</td>
</tr>
</tbody>
</table>

Key Terms

**Computer Simulation**
Artificially re-creating a random process under a specific set of conditions using a computer.

Introduction to Computer Simulation

Recall that simulation is a process of artificially re-creating a random process under a specific set of conditions. We can use computer simulation to help us determine whether an assumed value of a population parameter is plausible based on a calculated sample statistic. A computer can very quickly and easily generate multiple simulated sample statistics based on an assumed value of a population parameter and given a particular sample size. This section will demonstrate how to use some of the applets found at [http://www.rossmanchance.com/applets/index2021.html](http://www.rossmanchance.com/applets/index2021.html) to create simulated distributions.

Suppose I have a fair die. If my die is fair, then that would mean that it rolls the number “five” 1/6 of the time in the long run (1/6 = 0.167 = 16.7%). But, when rolling a die, we don’t necessarily expect to see exactly 1/6 of the rolls land on the number “five” (or on any specific value, for that matter). We expect some variation in the results. We expect that any sample of rolls will result in approximately 1/6 of the rolls landing on each of the numbers, one through six. But, how much variation is usual/expected vs. unusual/unexpected?

Suppose that I collect a sample by rolling my die 60 times, and it rolls the number five a total of 14 times in my sample. Since 14/60 = 0.233 = 23.3%, the number five was rolled more than 1/6 of the time. In 60 rolls, we expect to roll the number five about 10 times, but is 14 fives in 60 rolls unusually high for a fair die? How unusual is it to roll a five 14/60 = 23.3% of the time if this is a fair die?

In this scenario, the population is all rolls of the die, and the sample is the 60 times we observed the die being rolled. The parameter of interest is the proportion of times that the die rolls the number five in the long run. The value of the parameter, given that we are told that the die is fair, is \( \pi = \frac{1}{6} = 0.167 = 16.7\% \).

The statistic is the proportion of times that the die rolls the number five in 60 rolls. The statistic that we observed in the sample is \( \hat{p} = \frac{14}{60} \approx 0.233 = 23.3\% \).

We want to understand whether our statistic, \( \hat{p} = \frac{14}{60} \approx 0.233 = 23.3\% \), is a “usual/expected” sort of statistic, or if it is an “unusual/unexpected” sort of statistic. To understand this, we will examine what the
distribution of sample statistics would look like when the die is actually fair. If we roll a die 60 times, what proportion of fives would be usual/expected? The distribution of “potential” p-hat values will allow us to understand the variability we can expect among the sample statistics. It will show us how spread out the sample statistics are expected to be. Once we have that distribution of sample statistics, then we can compare our observed statistic to those found in the distribution of simulated sample statistics to see whether our observed statistic is usual/expected or unusual/unexpected.

In order to create the distribution of sample statistics, we could roll a fair die 60 times and record the proportion of times it rolls a five, and then roll it 60 times and record the statistic again, and then roll it 60 times and record the statistic again. If we did this many, many times (say, 100 or 1000 sets of 60 rolls), we would be able to get an understanding of what sort of sample statistics are usual and expected if the die is fair. We could then compare those sample statistics to our actual observed statistic of \( \hat{p} = \frac{14}{60} = 0.233 = 23.3\% \) to see if our observed statistic is usual/expected or unusual/unexpected.

Of course, rolling the fair die over and over in this manner would be incredibly time consuming. Instead, we will use a computer simulator in order to create the set of sample statistics that we would expect to see if the die is fair and it is rolled 60 times. In other words, we will use the computer simulator to create the chance model.

**Steps to Create Distribution of Simulated Sample Statistics for One Proportion**

1. Access the computer simulator, called the One proportion inference applet, by going to [http://www.rossmanchance.com/applets/2021/oneprop/OneProp.htm](http://www.rossmanchance.com/applets/2021/oneprop/OneProp.htm). Bookmark this applet since you’ll use it a lot!

2. Note that we are considering a scenario with a categorical variable (the variable is whether the die rolls the number five), and the statistic is the proportion of times the die rolls the number five. In general, when we consider a scenario involving a categorical variable we will use the One proportion inference applet.

3. Check the “Hide coins” box on the far right, and fill in the top three boxes as follows:

   - Probability of success (\( \pi \)): 0.167
   - Sample size (\( n \)): 60
   - Number of samples: 100

   “Probability of heads” will change to say “Probability of success (\( \pi \))” if you enter a value other than 0.5. This is the box where you need to enter the assumed value of the population parameter (which is 0.167 in this case). Note that you cannot enter the exact fraction 1/6; only a decimal value is accepted.

   “Number of tosses” will change to say “Sample size (\( n \))”. This is the box where you enter your sample size (which is 60 in this case).

   “Number of repetitions” will change to say “Number of samples”. This is the box where you enter the number of repetitions (usually 100 or 1000). We will do 100 repetitions this time. This is like 100 people each rolling the die 60 times, and then all of us come back together and share our results with each other. This will create 100 dots in the dot plot.

   Click the “Draw Samples” button to create the distribution.
You will now see a distribution of simulated sample statistics that displays 100 dots (because you input 100 as the “Number of samples”). This distribution can now be used to help us understand the sample statistics we should expect if the die is fair. **Notice that the snapshot of the distribution shown below will be slightly different than the distribution on your computer.** Everyone will get slightly different chance models when running a simulation using the applet (slightly different distributions of sample statistics), but they will all be extremely similar!

If you select “**Number of successes**”, the chance model will show you the 100 simulated “number of success” statistics that you should expect when rolling a fair die 60 times. For example, we see that there were 2 samples (2 dots) in which the fair die was rolled 60 times, and the die rolled a five 15 times out of 60 rolls. As another example, we see that there were 7 samples (7 dots) in which the fair die was rolled 60 times, and the die rolled a five 6 times out of 60 rolls.

When rolling the die 60 times (and then repeating that experiment 100 times total), my chance model shows that, on average, we roll a five 9.580 (about 10) times out of 60 rolls. The mean of this distribution is 9.580, and the standard deviation of this distribution (the standard error) is 2.882. Clicking the “**Summary Stats**” button displays the mean and standard deviation (SD) of the distribution of simulated sample statistics.

If you select “**Proportion of successes**”, the chance model will show you the 100 simulated “proportion of success” statistics that you should expect when rolling a fair die 60 times. For example, we see that there were 2 samples (2 dots) in which the fair die was rolled 60 times, and the die rolled a five 0.25 = 25% of the time (15 times out of 60 rolls). As another example, we see that there were 7 samples (7 dots) in which the fair die was rolled 60 times, and the die rolled a five 0.10 = 10% of the time (6 times out of 60 rolls).

When rolling the die 60 times (and then repeating that experiment 100 times total), my chance model shows that, on average, we roll a five 0.160=16.0% of the time. The mean of this distribution is 0.160, and the standard deviation of this distribution (the standard error) is 0.048. Clicking the “**Summary Stats**” button displays the mean and standard deviation (SD) of the distribution of simulated sample statistics.

If you select “**Number of successes**”, the chance model will show you the 100 simulated “number of success” statistics that you should expect when rolling a fair die 60 times. For example, we see that there were 2 samples (2 dots) in which the fair die was rolled 60 times, and the die rolled a five 15 times out of 60 rolls. As another example, we see that there were 7 samples (7 dots) in which the fair die was rolled 60 times, and the die rolled a five 6 times out of 60 rolls.

When rolling the die 60 times (and then repeating that experiment 100 times total), my chance model shows that, on average, we roll a five 9.580 (about 10) times out of 60 rolls. The mean of this distribution is 9.580, and the standard deviation of this distribution (the standard error) is 2.882. Clicking the “**Summary Stats**” button displays the mean and standard deviation (SD) of the distribution of simulated sample statistics.

If you select “**Proportion of successes**”, the chance model will show you the 100 simulated “proportion of success” statistics that you should expect when rolling a fair die 60 times. For example, we see that there were 2 samples (2 dots) in which the fair die was rolled 60 times, and the die rolled a five 0.25 = 25% of the time (15 times out of 60 rolls). As another example, we see that there were 7 samples (7 dots) in which the fair die was rolled 60 times, and the die rolled a five 0.10 = 10% of the time (6 times out of 60 rolls).

When rolling the die 60 times (and then repeating that experiment 100 times total), my chance model shows that, on average, we roll a five 0.160=16.0% of the time. The mean of this distribution is 0.160, and the standard deviation of this distribution (the standard error) is 0.048. Clicking the “**Summary Stats**” button displays the mean and standard deviation (SD) of the distribution of simulated sample statistics.
Recall that our goal is to determine whether our observed sample statistic is usual/expected or unusual/unexpected. Recall that we were previously told that we rolled the die 60 times and observed it roll the number five a total of 14 times, so that the observed sample statistic is \( \hat{p} = \frac{14}{60} = 0.233 = 23.3\% \).

We will now use the distribution of simulated sample statistics that we have created with the One proportion inference applet in order to determine how usual or unusual this observed sample statistic is from a fair die.

Using the simulated distribution that we created with the applet, we see that there are 9 dots out of a total of 100 dots which are at least as large as our observed sample statistic. That is, there are 9 dots out of a total of 100 dots where the die rolled a five 14 times or more in 60 rolls. This means that 9\% (9/100) of the sample statistics are like our observed sample statistic or higher. We could say that, if the die is fair, then we have a 9\% chance of observing a sample statistic at least as large as ours. A 9\% chance is higher than a 5\% chance, and so, according to the 5\% Rule, we would conclude that this is a “usual/expected” sort of statistic to get when the die is fair.

It turns out that the applet can determine this percentage for us, instead of forcing us to count dots. The instructions below detail how to use the applet to find this percentage. This feature of the applet will give us a quick way to determine whether our observed statistic is “usual” vs. “unusual” since it tells us the probability of getting a sample statistic at least as extreme as ours.
Steps to Determine the Percentage of Statistics at Least as Extreme as the Observed

1. Create the distribution of simulated sample statistics (see instructions above). Determine/decide whether you will use the “Number of successes” option or the “Proportion of successes” option. It is highly recommended that you always opt for “Proportion of successes” since the parameter and statistic are almost always given as proportions rather than counts. However, it is important that you understand how to determine the percentage no matter which option is used.

2. Enter the observed sample statistic in the “As extreme as” box. Be sure to enter the value appropriately to match with “Number of successes” or “Proportion of successes.”

   For example, if we observed 14 successes out of 60 trials ($\hat{p} = 14/60 = 0.233 = 23.3\%$), then we would enter 14 for the statistic if using “Number of successes” option.

   For example, if we observed 14 successes out of 60 trials ($\hat{p} = 14/60 = 0.233 = 23.3\%$), then we would enter 0.2333 for the statistic if using the “Proportion of successes” option.

3. Use the button to select either $\geq$ or $\leq$ based on which dots you want the applet to count.
   - The $\geq$ will count dots that are like the observed statistic or higher.
   - The $\leq$ will count dots that are like the observed statistic or lower.

   Note that our goal is to identify the proportion of statistics that are at least as extreme as the observed statistic, so that we can determine the percentage of the time we get a statistic like the observed or even more unusual than this. This means that you should count dots that are like the observed statistic or further away from the mean. In other words, we count dots that are like the observed statistic or further out into the tail of the distribution.

4. The applet will create a vertical red line located at the observed statistic. The dots that are like the observed statistic or more extreme will be turned red in the applet. The value of the observed statistic will be displayed at the bottom of the vertical line, and the top of the vertical line will display the proportion of dots that are like the observed statistic or more extreme. The proportion will also be displayed in red underneath the “As extreme as” area.

50
For example, in the die rolling example, \( \frac{9}{100} = 9\% \) of the dots are at least as large as our observed statistic. According to the 5\% Rule, our observed statistic is a usual/expected statistic to get if the die is fair.

**Example 1.4.1**

Candidate Johnson earned 48\% of the votes in an election. Ahmed randomly speaks with 75 people who voted in the election, and he finds that 33 of the people in his sample voted for Candidate Johnson. That is, \( \frac{33}{75} = 0.44 = 44\% \) of his sample voted for Johnson.

(a) Use the **One proportion inference** applet to create a distribution of 100 simulated sample statistics in order to determine the statistics that Ahmed should expect when sampling 75 voters, if 48\% of the population of voters voted for Candidate Johnson. Check the “Summary Stats” box so that the mean and standard deviation are displayed. Copy and paste a screenshot from the “Number of successes” option, as well as a screenshot from the “Proportion of successes” option.

(b) Interpret the contextual meaning of each dot in the dot plot.

(c) Use the applet to identify the percentage of simulated statistics that are at least as small as Ahmed’s observed statistic. Copy and paste a screenshot from the “Number of successes” option, as well as a screenshot from the “Proportion of successes” option. Also interpret, with a complete sentence in context, the meaning of the value in red found in the applet.

(d) Is Ahmed’s sample statistic unusual? Or is it a usual/expected statistic to get? Explain in context.

**Solution 1.4.1**

(a) The screenshots below show the correct values to enter for this example, and an example of the resulting distribution. Recall that everyone will get a slightly different distribution when using simulation.
Solution 1.4.1 (continued)

(b) Each dot in the dot plot represents the proportion of voters (or number of voters), out of 75 sampled voters, who voted for Candidate Johnson, given that 48% of all voters voted for Candidate Johnson. Each dot represents a simulated \( \hat{p} \). This simulated distribution resembles the generic distribution of sample p-hats:

(c) The screenshots below show the correct values to enter for this example in order to find the percentage of statistics that are like Ahmed’s or less. We find that \( \frac{28}{100} = 0.028 = 28\% \) of sample statistics are at least as small as Ahmed’s observed statistic. So, given that 48% of all voters voted for Candidate Johnson, we find that there is a 28% chance of getting a sample statistic like Ahmed observed or less (getting a sample statistic at least as low as 44% support for Candidate Johnson).

(d) Ahmed’s observed sample statistic, where he observed \( \hat{p} = \frac{33}{75} = 0.44 \) voting for Candidate Johnson is a usual/expected sort of statistic to get since there is a 28% chance of getting a statistic at least as low as 0.44 when sampling 75 voters. A 28% probability is much higher than 5%, and so \( \hat{p} = 0.44 \) is a usual/expected statistic according to the 5% Rule.
**Introduction to Distributions of Simulated Sample Statistics for a Mean**

In this section we will only explore the features of the distribution of simulated sample statistics for a mean when the distribution is already given. **We will wait until the next section** to learn the specific steps for creating such a distribution since that will require additional background information that we have yet to discuss.

Suppose that a soda company has machines that fill all cans of soda with an average (mean) of 12 ounces of soda. Juan collects a random sample of 38 cans, and his cans have an average (mean) of 11.92 ounces of soda. Juan is wondering whether his sample of soda cans has an unusually low average amount of soda in it, or if 11.92 ounces is a usual/expected mean to get. Below are some screen shots from a distribution of 100 simulated sample statistics for this scenario. This computer simulation creates little squares to represent each simulated statistic (instead of little dots).

When it comes to determining the percentage of simulated statistics that are at least as extreme as the observed statistic, we find that this applet has a similar procedure to the previously discussed applet. We need to select either $\leq$ or $\geq$, and we need to enter the observed sample statistic in the box for “Count Samples”. The computer applet will then count the squares that are at or below the observed sample statistic (Juan’s sample statistic was $\bar{x} = 11.92$) since Juan is interested in understanding whether his sample statistic is unusually low.

The “Less than $\leq$” option is selected because we want to count samples that are less than or equal to $\bar{x} = 11.92$.

We enter the sample statistic $\bar{x} = 11.92$ in the “Count Samples” box and click the “Count” button.

The applet counts 0 simulated sample statistics out of 100 total statistics that are less than or equal to the observed statistic. That is, 0% of the sample statistics are at least as low as the observed statistic. A sample mean of $\bar{x} = 11.92$ is an unusual statistic to get when the soda company is filling all cans with an average (mean) of 12 ounces.
Example 1.4.2

The average (mean) income for all households in the U.S. in 2019 was $89,930.70. Li collects a random sample of 52 U.S. households, and his sample had a mean income of $90,180.24 in the year 2019. Li would like to understand whether the sample he collected has a mean income that is unusually high, or if it is a usual/expected sort of mean to get. A distribution of 100 simulated sample statistics is given below for this scenario.

(a) Describe the population.
(b) Describe the parameter, give the symbol for the parameter, and give the value of the parameter.
(c) Describe the statistic, give the symbol for the statistic, and give the value of the statistic.
(d) Use the given distribution of 100 simulated sample statistics to identify the mean and standard deviation of the distribution.
(e) According to the 2 Standard Deviations Rule, is Li’s sample statistic “unusually high”? Explain.
(f) According to the 2 Standard Deviations Rule, when collecting a random sample of 52 households, it would be unusual to get a sample mean income that is less than __________ dollars, or a sample mean income that is more than __________ dollars.
(g) Use the given distribution of 100 simulated sample statistics to identify the proportion of statistics that are at least as high as Li’s observed statistic. Also, use that proportion to explain whether Li’s observed statistic is usual/expected or unusual/unexpected. Explain how you arrived at the conclusion about usual vs. unusual.

Solution 1.4.2

(a) The population is all households in the U.S. in 2019.
(b) The parameter is the average (mean) income of all households in the U.S. in 2019. The parameter value is usually unknown, but in this example we have been given the value of the parameter, which is $\mu = 89,930.70$.
(c) The statistic is the average (mean) income of the sample of 52 households in the U.S. in 2019. The value of the observed statistic is $\bar{x} = 90,180.24$.
(d) The mean of the distribution of simulated sample statistics is $89,886.263$ which is very close to the true value of the population parameter ($89,930.70$). It is slightly off since this is a simulation. Everyone who runs the computer simulation will get a slightly different distribution with a slightly different mean. The standard deviation of the distribution is $599.496$.

(continued on next page)
Solution 1.4.2 (continued)

(e) 2 standard deviations above the mean would be $89,886.263 + $599.496 + $599.496 = $91,085.26. We would consider sample statistics that are higher than $91,085.26 to be “unusually high” according to the 2 Standard Deviations Rule. Li’s observed statistic is $90,180.24. Since Li’s observed statistic is lower than $91,085.26, we would conclude that Li’s observed statistic is a usual/expected statistic to get.

(f) In part (e) we determined that $91,085.26 is the statistic that is 2 standard deviations above the mean. We now calculate $89,886.263 − $599.496 − $599.496 = $88,687.27 to find the statistic that is 2 standard deviations below the mean.

According to the 2 Standard Deviations Rule, when collecting a random sample of 52 households, it would be unusual to get a sample mean income that is less than $88,687.27 dollars, or a sample mean income that is more than $91,085.26 dollars.

The distribution of simulated sample statistics resembles the general distribution of sample x-bars, and we now understand that it can be labeled as follows:

(g) The screenshot below shows that 30 squares out of 100 total squares in the distribution are at least as high as Li’s observed statistic. This means that $\frac{30}{100} = 0.3 = 30\%$ of the simulated statistics are like Li’s statistic or higher. We could say that, when collecting a random sample of 52 households from a population that has a mean income of $89,886.263, there is a 30\% chance of getting a sample statistic like $90,180.24 or higher. Using the 5\% Rule, since there is a 30\% chance, which is higher than 5\%, we would conclude that $\bar{x} = 90,180.24$ is a usual/expected statistic to get.
Exercises

1) **(Multiple Choice)** When using the applets [One Proportion](#) (for categorical data) or [One Variable with Sampling](#) (for quantitative data) to generate a distribution of simulated statistics, the distribution you obtain will have a mean of approximately __________.
   (a) the value of the sample statistic
   (b) approximately half of the sample size
   (c) the number of repetitions/samples
   (d) the value of the population parameter

2) **(Multiple Choice)** Suppose a researcher is testing to see whether a basketball player can make free throws at a rate higher than the NBA average of 75%. The player is tested by shooting 10 free throws and makes 8 of them. Suppose a computer applet is used to conduct an appropriate simulation with 1000 repetitions, and a distribution of simulated statistics is created. What does this distribution represent?
   (a) Repeated results if the player makes 80% of his free throws in the long run.
   (b) Repeated results if the player makes 75% of his free throws in the long run.
   (c) Repeated results if the player makes more than 75% of his free throws in the long run.
   (d) Repeated results if the player makes more than 80% of his free throws in the long run.

3) A common test for extrasensory perception (ESP) has the researcher actively think about a number from 1 to 5 as the subject attempts to determine which number the researcher chose by “reading the researcher’s mind”. Suppose we want to determine whether our friend, Gary, has ESP. We conduct 26 of the previously described tests and find that Gary selects the correct number 9 times and fails to identify the correct number 17 times.
   (a) Describe the parameter of interest. What symbol should be used to represent the parameter?
   (b) What is the value of the parameter if Gary is just guessing at numbers randomly?
   (c) If Gary is not just guessing (something else is affecting his choices), would you expect the parameter value to be larger or smaller than the value stated in (b)? Explain.
   (d) What is the sample size in the study? Include the correct symbol in your answer.
   (e) Describe the statistic. Give the value of the sample statistic using the correct symbol.

4) You were sold a six-sided die that is supposed to land on the numbers 1 and 2 more frequently than they should occur with a fair die. To determine whether the die is weighted to favor values of 1 and 2 you roll it a total of 65 times and recorded a result of 1 or 2 on 22 of the rolls, while 43 rolls resulted in a 3, 4, 5, or 6.
   (a) Describe the parameter of interest. What symbol should be used to represent the parameter?
   (b) What is the value of the parameter if the die is fair (not weighted)?
   (c) If the die is weighted to favor values of 1 and 2, would you expect the parameter value to be larger or smaller than the value stated in (b)? Explain.
   (d) What is the sample size in the study? Include the correct symbol in your answer.
   (e) Describe the statistic. Give the value of the sample statistic using the correct symbol.
5) Josh is a foosball player who is working on his trick shots. While practicing, he gets 15 out of 20, or 75%, of his trick shots in the goal. Josh wants to know whether there is evidence that he is more likely than not to make his trick shots in the long run.

(a) Describe the statistic. Give the value of the sample statistic using the correct symbol.
(b) Describe the parameter of interest and give the correct symbol for the parameter.
(c) What is the value of the parameter if Josh is just as likely to miss as he is to make his trick shot?
(d) Suppose you want to create a distribution of 100 simulated sample statistics. A screenshot from the One proportion inference applet, including a distribution of simulated sample statistics, is shown below. Fill in the blanks that were used to generate the distribution.

(e) (Fill in the blanks) According to the 2 Standard Deviations Rule, if Josh is just as likely to make as he is to miss his trick shots, then out of 20 shots it would be unusual for his proportion of made shots to be less than __________ or more than __________. Show how you find the answers.

(f) Use the distribution of 100 simulated sample statistics shown in part (d) to identify the proportion of statistics that are at least as large as Josh’s observed statistic. Explain how you find the proportion, then use your answer to fill in the blank in the statement below.

Assuming Josh is just as likely to make his trick shots as to miss them, there is a __________% chance of him making 15 or more trick shots out of 20 attempts.
6) The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all men between the ages of 20 and 29 to be 5.8 feet (5 feet 9.5 inches) with a standard deviation of 0.35 feet (4.2 inches). You want to determine whether the average height of men in this age group is higher among Dutchess County residents. You survey a random sample of 74 Dutchess County residents between the ages of 20 and 29 that identified as men and measured an average height of 5.91 feet (5 feet 10.92 inches).

(a) Describe the parameter of interest in the context of the study and determine which symbol should be used to represent the parameter.

(b) Describe the statistic. Give the value of the sample statistic using the correct symbol.

(c) Suppose you want to create a distribution of 100 simulated sample statistics. A screenshot from the One Variable with Sampling applet, including a distribution of simulated sample statistics, is shown below. Fill in the blanks that were used to generate the distribution.

(d) (Fill in the blanks) According to the 2 Standard Deviations Rule, if the average height of all men ages 20-29 in Dutchess County is 5.8 feet, it would be unusual to collect a random sample of 74 men ages 20-29 from Dutchess County and find a sample mean height that is less than _______ or more than _______. Show how you find the answers.

(e) Use the distribution of 100 simulated sample statistics shown in part (c) to identify the proportion of statistics that are at least as extreme as the observed sample statistic. Explain how you find the proportion, then use your answer to fill in the blanks in the statement below.

Assuming the average height of all Dutchess County men ages 20-29 is 5.8 feet, the probability of obtaining a random sample of 74 Dutchess County men ages 20-29 with an average height of __________ or higher is __________.
7) It has been claimed that in some sports there might be a competitive advantage for teams wearing blue uniforms. You want to test this with the sport of competitive swimming. You conduct a study using 36 separate races where 3 competitive swimmers are chosen at random and assigned either a blue, red, or green swimsuit for each race. The swimmer wearing a blue swimsuit won in 18 of the 36 races.

(a) Describe the statistic. Give the value of the sample statistic using the correct symbol.

(b) Describe the parameter of interest in the context of the study and state which symbol should be used to represent the parameter.

(c) What is the value of the parameter if wearing a blue swimsuit does not provide a competitive advantage (if each of the three swimsuit colors is equally likely to win)?

(d) Does the sample statistic prove that wearing blue provides a competitive advantage? Explain using complete sentences.

(e) Suppose you want to create a distribution of 300 simulated sample statistics. A screenshot from the One proportion inference applet, including a distribution of simulated sample statistics, is shown below. Fill in the blanks that were used to generate the distribution.

(f) (Fill in the blanks) The distribution above has a mean of 0.329 and a standard deviation of 0.081. According to the 2 Standard Deviations Rule, when sampling 36 competitive swimming races where a blue swimsuit does not provide a competitive advantage, it would be unusual for the wearer of the blue swimsuit to win fewer than ______% or more than ______% of their races. Show how you find the answers.

(g) Use the distribution of 300 simulated sample statistics shown in part (e) to identify the proportion of statistics that are at least as extreme as the observed statistic. Explain how you find the proportion, then use your answer to fill in the blanks in the statement below.

Assuming blue swimsuits do not provide a competitive advantage in a swimming race, the probability of the competitor wearing the blue swimsuit winning at least 18 out of 36 races is _________. This means there is evidence to suggest our sample statistic is ________ (usual or unusual).
8) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that 2 out of every 3 people prefer their brand of cola. A random sample of 140 people is taken, and 85 of them (60.7%) choose Pepsi as their preferred cola.

(a) What is the assumed/claimed value of the population parameter? Also identify the correct symbol for the parameter.

(b) What is the value of the observed statistic? Also identify the correct symbol for the statistic.

(c) Use the One proportion inference applet to create a distribution of 1000 simulated sample statistics for this situation. Check the “Summary Stats” box to display the mean and standard deviation of the distribution. Take a screenshot and paste it into your solution here. Your screenshot should display the values you entered into the applet as well as the distribution generated.

(d) According the 2 Standard Deviations Rule, what conclusion should be made regarding the claimed value of the parameter? Write a complete sentence in context to state the conclusion.

9) The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all women ages 20-29 to be 5.375 feet (5 feet 4.5 inches) with a standard deviation of 0.283 feet (3.4 inches). You want to determine whether the average height of Dutchess County women ages 20-29 is different than the national average. You survey a random sample of 36 Dutchess County residents ages 20-29 that identify as women and find an average height of 5.485 feet (5 feet 5.8 inches).

(a) What is the assumed/claimed value of the population parameter? Also identify the correct symbol for the parameter.

(b) What is the value of the observed statistic? Also identify the correct symbol for the statistic.

(c) A distribution of 1000 simulated sample statistics is shown below. According to the 2 Standard Deviations Rule, what conclusion should be made regarding the claimed value of the parameter? Write a complete sentence in context to state the conclusion.
10) A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest was for workers who live in New York City, who traveled an average of 35.9 minutes each day (or a total of 89.4 hours travel time per year) with a standard deviation of 12.5 minutes. You live relatively close to the city and know that many of your neighbors work there, so you decide to investigate whether people in your neighborhood have a longer average commute time than the average worker who lives in New York City. You survey a total of 38 commuters in your neighborhood and find an average commute time of 39.7 minutes, with a standard deviation of 6.1 minutes.

(a) What is the assumed/claimed value of the population parameter? Also identify the correct symbol for the parameter.

(b) What is the value of the observed statistic? Also identify the correct symbol for the statistic.

(c) A distribution of 1000 simulated sample statistics is shown below. According to the 2 Standard Deviations Rule, what conclusion should be made regarding the claimed value of the parameter? Write a complete sentence in context to state the conclusion.

11) A study was conducted by the National Endowment for the Arts to describe music preferences in the Hudson Valley. The 2010 report describes relationships between key demographic characteristics and music preferences. Researchers found that 24% of adults preferred the genre of Rock/Heavy Metal at that time. Recently, a random sample of 715 Hudson Valley adults was surveyed with 150 indicating they prefer Rock/Heavy Metal. You want to determine whether the proportion of Hudson Valley residents that prefer Rock/Heavy Metal has changed since the 2010 report.

(a) What is the assumed/claimed value of the population parameter? Also identify the correct symbol for the parameter.

(b) What is the value of the observed statistic? Also identify the correct symbol for the statistic.

(c) Use the One proportion inference applet to create a distribution of 1000 simulated sample statistics for this situation. Check the “Summary Stats” box to display the mean and standard deviation of the distribution. Take a screenshot and paste it into your solution here. Your screenshot should display the values you entered into the applet as well as the distribution generated.

(d) According to the 2 Standard Deviations Rule, what conclusion should be made regarding the claimed value of the parameter? Write a complete sentence in context to state the conclusion.
Section 1.5 | Additional Details on Quantitative Data

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fluently use the following vocabulary and symbols:</td>
</tr>
<tr>
<td>• Sample standard deviation, $s$</td>
</tr>
<tr>
<td>• Population standard deviation, $\sigma$ (sigma)</td>
</tr>
<tr>
<td>• Standard deviation of the distribution of sample statistics, standard error</td>
</tr>
<tr>
<td>• Median</td>
</tr>
<tr>
<td>• Outlier</td>
</tr>
<tr>
<td>• Left-skewed distribution</td>
</tr>
<tr>
<td>• Right-skewed distribution</td>
</tr>
<tr>
<td>• Estimate the standard deviation of a simple data set by estimating the average distance that the data values are from them mean.</td>
</tr>
<tr>
<td>• Understand the relationship between the mean and median of a skewed distribution.</td>
</tr>
<tr>
<td>• Understand the effect of outliers on the mean, median, and skew of a distribution.</td>
</tr>
<tr>
<td>• Use a computer applet to create a distribution of simulated sample means.</td>
</tr>
</tbody>
</table>

Key Terms

**Standard deviation:**
In simple terms, the standard deviation of a set of quantitative data is the average distance of the data values from the mean of the data set. Standard deviation is a measure of variability among the data values.

**Sample standard deviation, $s$:**
The standard deviation of a set of sample data.

**Population standard deviation, $\sigma$ (sigma):**
The standard deviation of an entire population of data.

**Median:**
A measure of center in a data set. It is the middle value, the value that has exactly 50% of data above it or equal to it, and 50% of the data is below it or equal to it. The median is not very effected by outliers.

**Outlier:**
Data that is markedly higher or lower than the bulk of the data.

**Left-skewed distribution:**
A distribution that has one, or a small number, of outliers on the left side. The mean of a left-skewed distribution is “pulled” towards the outlier(s) on the left, and so the mean of a left-skewed distribution is on the left side of (less than) the median of that distribution.

**Right-skewed distribution:**
A distribution that has one, or a small number, of outliers on the right side. The mean of a right-skewed distribution is “pulled” towards the outlier(s) on the right, and so the mean of a right-skewed distribution is on the right side of (greater than) the median of that distribution.
Introduction to Standard Deviation

The standard deviation of a set of quantitative data is a number that tells us the average amount of “spread” in that data set. While the mean is a measure of where the center of a data set is, the standard deviation is a measure of the variability among the data values. A larger standard deviation indicates that the data values are more spread out (vary more), and a smaller standard deviation indicates that the data values are closer together (vary less). In simple terms, the standard deviation of a data set tells us the average distance that the values in the data set are away from the mean. The technical definition of the standard deviation, which is the value given in calculators, is more complicated than this but we won’t need to use that formal calculation in this class.

- If the quantitative data set we are working with is a set of sample data, then the sample standard deviation is denoted $s$.
- If the quantitative data set we are working with is the full set of population data, then the population standard deviation is denoted $\sigma$ (sigma).
- If we are working with a distribution of sample statistics, then the standard deviation of that distribution of sample statistics has a special name: the standard error. The standard deviation of the distribution of sample statistics is called the standard error.

Suppose we consider annual employee salaries at three different companies as given below:

<table>
<thead>
<tr>
<th>Company A</th>
<th>Company B</th>
<th>Company C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$150,000</td>
<td>$130,000</td>
<td>$500</td>
</tr>
<tr>
<td>$150,000</td>
<td>$140,000</td>
<td>$500</td>
</tr>
<tr>
<td>$150,000</td>
<td>$150,000</td>
<td>$150,000</td>
</tr>
<tr>
<td>$150,000</td>
<td>$160,000</td>
<td>$299,500</td>
</tr>
<tr>
<td>$150,000</td>
<td>$170,000</td>
<td>$299,500</td>
</tr>
</tbody>
</table>

If we calculate the mean salary at each company, we find that each company has a mean salary of $150,000. In other words, the center of each data set is the same. However, the variability in the salaries is not the same for each company. The standard deviation is a number that describes the average distance the data values are from the mean. In other words, the standard deviation tells us how far away from the mean each data value is on average.

- The standard deviation of the sample salaries at Company A is $s = 0$. This tells us that on average, the salaries at Company A are $0 away from the mean salary. This means that the data has NO spread away from the mean, or that the data values do not vary. We can find that standard deviation by just looking at the data set since we see that there is 0 distance from each data value to the mean. There is no variability at all in this data, which means the standard deviation is $0$. But, if necessary, we could also copy and paste the data into the Descriptive Statistics applet to find the values of the mean and standard deviation:
• The standard deviation of the sample of salaries at Company B is $s = 15,811.4$. On average, the salaries at Company B are $15,811.40$ away from the mean salary. This tells us that the data has SOME variability. We can find the mean and standard deviation values by copying and pasting the data set into the Descriptive Statistics applet:

![Descriptive Statistics applet for Company B](image)

The standard deviation value for the Company B data set could be estimated as about $12,000$ without the use of the applet by finding the distance that each data value is from the mean, and then averaging those distances. Note that the distances are all positive numbers, so the data values $130,000$ and $170,000$ are both the same distance from the mean (20,000), even though $130,000$ is below the mean and $170,000$ is above the mean.

<table>
<thead>
<tr>
<th>Original data value</th>
<th>Distance the data value is from the mean (mean is 150,000)</th>
<th>Estimate of the standard deviation (average of those distances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>130,000</td>
<td>20,000</td>
<td>$\frac{20,000 + 10,000 + 0 + 10,000 + 20,000}{5} = 12,000$</td>
</tr>
<tr>
<td>140,000</td>
<td>10,000</td>
<td></td>
</tr>
<tr>
<td>150,000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>160,000</td>
<td>10,000</td>
<td></td>
</tr>
<tr>
<td>170,000</td>
<td>20,000</td>
<td></td>
</tr>
</tbody>
</table>

So, although the actual standard deviation of Company B is $15,811.4$, we could estimate it as approximately $12,000$ without needing to use the applet. Our estimate is not perfect, but it gives us a quick sense of the variability in the salaries without having to resort to using a complicated statistical formula or having to use technology like the applet.

• The standard deviation of the salaries at Company C is $s = 149,500$. On average, the salaries at Company C are $149,500$ away from the mean salary. This tells us that the data has a LOT of variability! We can find the mean and standard deviation values by copying and pasting the data set into the Descriptive Statistics applet:

![Descriptive Statistics applet for Company C](image)

The standard deviation value for the Company C data set could be estimated as about $119,600$ without the use of the applet by finding the distance that each data value is from the mean, and then averaging those distances:
So, although the actual standard deviation of Company C is $149,500, we could estimate it as approximately $119,600 without needing to use the applet. Our estimate is not perfect, but it gives us a quick sense of the amount of variability in the salaries without having to resort to using a statistical formula or having to use technology like the applet.

<table>
<thead>
<tr>
<th>Original data value</th>
<th>Distance the data value is from the mean (mean is 150,000)</th>
<th>Estimate of the standard deviation (average of those distances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>149,500</td>
<td>149,500 + 149,500 + 0 + 149,500 + 149,500</td>
</tr>
<tr>
<td>500</td>
<td>149,500</td>
<td>5</td>
</tr>
<tr>
<td>150,000</td>
<td>0</td>
<td>= 598,000</td>
</tr>
<tr>
<td>299,500</td>
<td>149,500</td>
<td>5</td>
</tr>
<tr>
<td>299,500</td>
<td>149,500</td>
<td>= 119,600</td>
</tr>
</tbody>
</table>

Example 1.5.1

Suppose Bob conducts a study where he randomly selects 200 people at the mall and finds out their annual incomes. Suppose Tina conducts a study where she randomly selects 200 vascular surgeons in New York City and finds out their annual incomes.

(a) Which study would most likely have the higher sample mean, Bob's sample for the people at the mall, or Tina's sample for the vascular surgeons? Explain.

(b) Which data set would most likely have the higher standard deviation, Bob's sample of people from the mall, or Tina's sample of vascular surgeons? Explain.

Solution 1.5.1

(a) Tina's sample of vascular surgeons would most likely have the higher mean because vascular surgeons make a lot of money. If you pick random people at the mall, you would likely get some people who have small incomes, some people who have middle-range incomes, and some people who have high incomes. The average income for the people from the mall would most likely be a "middle range" income value. The mean vascular surgeon income would be high because all the vascular surgeon incomes are high.

(b) Bob's sample of people from the mall would most likely have the higher standard deviation. People randomly picked at the mall would have a wide variety of incomes: some low, some middle, and some high. Vascular surgeons all make somewhat similar salaries. So, there is low variability in the vascular surgeons’ salaries.
Mean, Median, and the Impact of Outliers

It is also important to understand that even though statistics texts (such as this one) focus a great deal of time on the mean of the data, it is sometimes the case that the “average” or “typical value” of a data set is better represented by the median rather than the mean.

It may be the median, and not the mean, that should be used when trying to get a sense of “average”.

The median, or middle value, of a set of quantitative data is the value for which 50% of the data is greater than or equal to that value, and 50% of the data is less than or equal to that value. An outlier is a data value that is markedly higher or lower than most of the data. One outlier, or a small number of outliers, can have a dramatic impact on the mean of the data set, but they will have little to no impact on the median of that data set. Let’s look at a simple example to understand the mathematics of this.

Example 1.5.2

Use the data set below to answer the questions.

100 100 100 100 100 100 100 100 100 5,000,000

(a) Does this data set have any outliers? If so, identify the outlier(s).

(b) Find the mean of the original data set. Then remove the outlier from the data set and re-calculate the mean without the outlier.

(c) Find the median of the original data set. Then remove the outlier from the data set and re-calculate the median without the outlier.

(d) Summarize the impact on the mean and median when the outlier is removed from the data set.

Solution 1.5.2

(a) There are ten data values, and one of them is an outlier. The outlier data value is the value 5,000,000 since that value is markedly higher than all the other values in the data set.

(b) The mean of the original data set is 500,090.

\[ \text{mean} = \frac{100 \times 9 + 5,000,000}{10} = \frac{900 + 5,000,000}{10} = \frac{5,000,900}{10} = 500,090 \]

If we remove the outlier value, then the data set has nine values that are all 100. The mean of the new data set is 100.

(c) The median of the original data set is 100. Half (50%) of the data values are bigger than or equal to 100, and half (50%) of the data values are less than or equal to 100. If we remove the outlier value, then the data set has nine values that are all 100. The median of that new data set is still 100 since 50% of the data values are bigger than or equal to 100, and 50% of the data values are less than or equal to 100.

(d) In the original data set, the mean is 500,090 and the median is 100. If we remove the very large outlier value, then the mean is 100 and the median is 100. The high outlier value “pulled” the mean higher but had no effect on the median.

In the “real world,” it is important to keep in mind that outliers can have a dramatic impact on the mean while having little to no impact on the median. One or more extreme data values can “pull” the mean value towards the extreme which can give us a “skewed” understanding of “average”. For that reason, when outliers are present it is beneficial to use the median to give us a sense of the typical data value instead of using the mean. Remember, the median isn’t going to be impacted dramatically by a few extreme outlier values, so the median can give us a better sense of the “average” data value when outliers are present.
Left-Skewed and Right-Skewed Distributions

Recall that a normal distribution is a distribution of data that is symmetrical, bell-shaped, and centered around the mean of the data. While it is true that textbooks on introductory statistics (such as this one) tend to focus a great deal of time on the normal distribution, it is important for students to understand that there are a great many shapes that a distribution of data can take on.

Not all data are distributed as a normal distribution!

Data can have many other distribution shapes besides the bell shape of a normal distribution. Two other common distribution shapes are a left-skewed distribution and a right-skewed distribution, as shown in the table below.

Recall from the previous page that outlier data values will “pull” the mean towards those extreme values, but the median won’t be impacted by outliers much, if at all. Even if a small number of data values are extremely high or low, the median will remain “in the middle”, which gives us a better sense of “average” than we would find from the mean. Therefore, the median can gives a better sense of “typical” when a distribution is skewed.

The example on the previous page, with nine data values equal to 100 and one data value equal to 5,000,000, was an example of a right-skewed distribution. The mean was “pulled” towards the extremely high outlier value while the median was not impacted by the outlier value:
Company A and B and C each employ 9 people. The salaries of the employees are given below.

<table>
<thead>
<tr>
<th>Company A</th>
<th>Company B</th>
<th>Company C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$110,000</td>
<td>$110,000</td>
<td>$50</td>
</tr>
<tr>
<td>$140,000</td>
<td>$140,000</td>
<td>$140,000</td>
</tr>
<tr>
<td>$140,000</td>
<td>$140,000</td>
<td>$140,000</td>
</tr>
<tr>
<td>$150,000</td>
<td>$150,000</td>
<td>$150,000</td>
</tr>
<tr>
<td>$150,000</td>
<td>$150,000</td>
<td>$150,000</td>
</tr>
<tr>
<td>$160,000</td>
<td>$160,000</td>
<td>$160,000</td>
</tr>
<tr>
<td>$160,000</td>
<td>$160,000</td>
<td>$160,000</td>
</tr>
<tr>
<td>$200,000</td>
<td>$1,000,000</td>
<td>$200,000</td>
</tr>
</tbody>
</table>

(a) Identify the shape each company’s salary distribution would most closely resemble: normal or left-skewed or right-skewed.

(b) Identify any outliers in the data set.

(c) Without doing any calculations, for each company, identify whether the mean salary would be equal to, or less than, or greater than the median salary.

(d) Use the **Descriptive Statistics** applet to identify the mean salary and the median salary for each company. Verify that your answers to parts (a) through (c) are correct.

**Solution 1.5.3**

(a) The salaries at Company A have a roughly normal distribution. The salaries at Company B have a right-skewed distribution because of the high outlier value of 1,000,000. The salaries at Company C have a left-skewed distribution because of the low outlier value of 50.

(b) Company A has no outliers. Company B has an extremely high outlier of 1,000,000. Company C has an extremely low outlier of 50.

(c) Company A would have a mean and median that would be approximately equal since it is a roughly normal distribution. Company B would have a mean that is higher than the median since it is a right-skewed distribution where the mean is “pulled” to the right towards the extremely high outlier. Company C would have a mean that is lower than the median since it is a left-skewed distribution where the mean is “pulled” to the left towards the extremely low outlier.

(d) Below are the distributions from the **Descriptive Statistics** applet, which show the distribution shapes as well as the means and medians for each company. This verifies that our answers above are correct.
Creating a Distribution of Simulated Sample Means using the Applet

Thus far in the text, students have been expected to use the One proportion inference applet to create distributions of simulated sample proportions for scenarios involving \( \pi \) and \( \hat{p} \) (when the variable is categorical). But, so far in the text we have provided you with the distribution of simulated sample means for problems involving \( \mu \) and \( \bar{x} \) (when the variable is quantitative). At this time, we will describe how to use the One Variable with Sampling applet to create the distribution of simulated sample means.

Steps to creating a distribution of simulated sample means using the One Variable with Sampling applet:

1. Open the “Sampling from a Finite Population/Model/Bootstrap” applet.

   ![Sampling from a Finite Population/Model/Bootstrap applet](image)

2. Select the “Population model” link on the bottom of the page.

   ![One Variable with Sampling applet](image)

(continued on next page)
3. Check the “Show Sampling Options” box.

4. Enter the information (mean, standard deviation, and shape) about the population, enter the sample size, and enter the number of sample statistics to generate. Then click the “Draw Samples” button.

Direct link to the applet above:  

Note that the population model applet requires us to identify the known or assumed population mean, \( \mu \), the known or assumed population standard deviation, \( \sigma \), as well as the known or assumed shape of the population distribution. In most cases, these three things are each generally unknown, and so we are essentially making assumptions or educated guesses about these things.

If you don’t know the value of the population standard deviation, \( \sigma \), then it is recommended that you enter the sample standard deviation, \( s \). The sample standard deviation is the only information we would have, and so it is our best guess as to the value of the unknown population standard deviation.

Important Note about the Applet for Simulating Sample Means

The distribution of simulated sample means can change a great deal if you modify the assumed population standard deviation, \( \sigma \), and/or if you modify the assumed shape of the population distribution.

If you don’t know the shape of the population distribution, then you can make an educated guess based on the context of the data. Sometimes we can logically make a good guess about how the data in the population would likely be distributed.
A cookie company claims that all of their cookies have an average (mean) of 24 chocolate chips per cookie. I randomly sample 40 cookies, and find that my sample of cookies has an average (mean) of 22.1 chocolate chips per cookie, with a sample standard deviation of 3.8 chips.

(a) Assign appropriate symbols to the numbers 24, 40, 22.1, and 3.8.

24: ________ 40: ________ 22.1: ________ 3.8: ________

(b) What is the claimed value of the population parameter, and what is the alternative range of values that the parameter could be if this claim were untrue?

(c) Create a distribution of simulated sample statistics and find the percentage of simulated statistics that are at least as extreme as the observed statistic.

(d) Make a conclusion about the company's claim based on the answer from part (c).

Solution 1.5.4

(a) \( \mu = 24 \) is the assumed population mean (the mean number of chips per cookie among all cookies).
\( n = 40 \) is the sample size (40 cookies were sampled).
\( \bar{x} = 22.1 \) is the sample mean (the mean number of chips per cookie in the sample of 40 cookies).
\( s = 3.8 \) is the sample standard deviation (the standard deviation for the sample of 40 cookies).

(b) The claimed value of the population parameter is \( \mu = 24 \), meaning that the average number of chips per cookie for ALL cookies made by the company is 24 chips per cookie. If this is not true, then it must be the case that the true value of the population mean is either lower than 24 or higher than 24. In other words, it would mean that the mean number of chips per cookie for ALL cookies is not 24 chips per cookie. In symbols, \( \mu \neq 24 \).

(continued on next page)
Example 1.5.4 continued

Solution 1.5.4 continued

(c) Open the Sampling from a Finite Population applet and click on the “Population model” link at the bottom of the page. The population mean is 24. The population standard deviation is unknown, but our best guess for this value is the sample standard deviation, which is 3.8. We don’t know the shape of the population standard deviation, but it is reasonable to think that the population of all cookies would have a normally distributed number of chips per cookie. So, we enter “Normal” for the population shape.

Check the “Show Sampling Options” box. We generate 1000 sample statistics, and the sample size is 40.

![Sampling from a Finite Population Applet](image)

To find the percentage of simulated sample statistics that are like our observed statistic or more extreme, we select the Count Samples “Beyond” option. This coincides with the fact that the true value of the parameter may be more or less than 24.

We need to count sample statistics that are at least as extreme as our sample statistic. In other words, we are counting statistics that are at least as extreme as \( \bar{x} = 22.1 \). We get an answer of 5 statistics out of a total of 1000 simulated statistics, which is 5/1000 = 0.005 = 0.5% of the simulated statistics are like the observed statistic or more extreme.

(d) If it is true that, as the company claims, there is an average of \( \mu = 24 \) chips per cookie in the population of all cookies, then the probability of getting a sample statistic at least as extreme as \( \bar{x} = 22.1 \) is 0.005 = 0.5%, which is less than 5%. According to the 5% Rule, the observed statistic is unusual/unexpected if the company’s claim is correct. We conclude that the mean number of chips per cookie in the population of all cookies is not 24 chips per cookie.
Section 1.5 | Additional Details on Quantitative Data

Exercises

1) I create a 5-question quiz on solving algebra equations. I gather a random sample of 100 students from the DCC cafeteria and have them take the quiz. I then gather a different random sample of 100 students from the afternoon sections of Calculus and have them take the quiz.
   a) Which group’s quiz scores do you expect to have the higher mean? Explain.
   b) Which group’s quiz scores do you expect to have the higher standard deviation? Explain.

2) My morning section of Introductory Statistics takes a quiz, and the standard deviation of their scores is 2. My evening section of Introductory Statistics takes a quiz, and the standard deviation of their scores is 10.
   a) Can you tell which group did better on the quiz on average? Explain.
   b) What (if anything) does the standard deviation tell us about how the students in each class are performing?

3) Acme has ten employees. Nine employees are line workers, and the tenth employee is the CEO. Below are the salaries of the ten employees:

<table>
<thead>
<tr>
<th>Salary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$42,000</td>
<td>$39,000</td>
</tr>
<tr>
<td>$40,500</td>
<td>$43,200</td>
</tr>
<tr>
<td>$40,570</td>
<td>$375,000</td>
</tr>
<tr>
<td>$39,750</td>
<td>$41,800</td>
</tr>
<tr>
<td>$42,000</td>
<td>$42,000</td>
</tr>
</tbody>
</table>

   (a) Find the mean and median salary. Show any work/calculations, or provide a screenshot if you use the Descriptive Statistics applet.
   (b) Are there any outliers? If so, what are their values?
   (c) Calculate the mean and median salary if the CEO's salary of $375,000 is NOT included in the salary list.
   (d) Write a rule summarizing how outliers tend to affect the mean and median of a data set.

4) Bob is moving to a new town. He looks online and learns that the mean cost of two-bedroom houses in the town is $435,000. He also learns that the median cost of a two-bedroom house in the town is $185,000. Bob is not sure which value he should use to help him anticipate the typical two-bedroom house cost in the area. Write to Bob and explain why it would make more sense for him to use the median. Be sure to include a statement that explains why the mean housing cost is so much higher than the median.

5) A census of the entire population concludes that in the year 2019 the average (mean) household income was $57,340.85 with a standard deviation of $4,598.35. A research company wants to know whether the mean household income in 2021 is lower than the average found in 2019, or whether the mean household income is the same in 2021 as it was in 2019. They randomly sample 258 households and their sample has a mean household income of $55,470.93 with a standard deviation of $5,650.42.
   (a) Identify the appropriate symbols for the numbers 57340.85, 4598.25, 258, 55470.93, and 5650.42.
   (b) What is the claimed value of the population parameter, and what is the alternative range of values that the parameter could be if this claim were untrue?
   (c) Use the Sampling from a Finite Population applet to create a distribution of simulated sample statistics and find the percentage of simulated statistics that are at least as extreme as the observed statistic. Copy and paste a screenshot from the applet that shows what you entered into the applet and also shows the resulting distribution.
   (d) Based on the answer from part (c), state a conclusion about the average household income in 2021.
An aluminum company has, for many years, created soda cans which have weights that are normally distributed with a mean weight of 0.5 ounces and a standard deviation of 0.03 ounces. The company purchases some new equipment. They collect a sample of 65 cans produced with the new equipment, and they find that these cans have a mean weight of 0.54 ounces with a standard deviation of 0.2 ounces. They want to know whether the new equipment is producing cans that still have a mean weight of 0.5 ounces, or whether the new equipment is actually creating cans with a larger weight on average.

(a) Identify the appropriate symbols for the numbers 0.5, 0.03, 65, 0.54, and 0.2.

(b) What is the claimed value of the population parameter, and what is the alternative range of values that the parameter could be if this claim were untrue?

(c) Use the Sampling from a Finite Population applet to create a distribution of simulated sample statistics and find the percentage of simulated statistics that are at least as large as the observed statistic. Copy and paste a screenshot from the applet that shows what you entered into the applet and also shows the resulting distribution.

(d) Based on the answer from part (c), state a conclusion about the mean weight of the cans produced by the new equipment.
Chapter 2 | Introduction to Hypothesis Testing

Contents of Chapter 2

Section 2.1 Introduction to Hypothesis Testing

Section 2.2 Introduction to P-value and Conclusions in a Hypothesis Test

Section 2.3 Standardized Statistics and Conclusions in a Hypothesis Test

Section 2.4 Strength of Evidence in a Conclusion
## Section 2.1 | Introduction to Hypothesis Testing

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fluently use the following vocabulary and symbols:</td>
</tr>
<tr>
<td>o Inferential statistics</td>
</tr>
<tr>
<td>o Research question</td>
</tr>
<tr>
<td>o Hypothesis test</td>
</tr>
<tr>
<td>o Null Hypothesis, $H_0$:</td>
</tr>
<tr>
<td>o Alternative hypothesis $H_A$:</td>
</tr>
<tr>
<td>o One-sided test</td>
</tr>
<tr>
<td>o Two-sided test</td>
</tr>
<tr>
<td>o Representative sampling method</td>
</tr>
<tr>
<td>o Biased sampling method</td>
</tr>
<tr>
<td>o Random sample</td>
</tr>
<tr>
<td>• Understand that the purpose of inferential statistics is to use observed, known, sample statistics to make evidence-based conclusions (inferences) about the unknown population parameters.</td>
</tr>
<tr>
<td>• Create null and alternative hypotheses for both $\pi$ and $\mu$ in both words and symbols.</td>
</tr>
<tr>
<td>• Differentiate between random sampling methods and non-random sampling methods.</td>
</tr>
<tr>
<td>• Understand the difference between a representative sampling method and a biased sampling method, and the reason for using a representative sampling method.</td>
</tr>
<tr>
<td>• Understand the purpose of collecting a random sample.</td>
</tr>
<tr>
<td>• Conduct the first 3 steps of a hypothesis test, and then use the 2 Standard Deviations Method to identify whether a sample statistic would be considered usual or unusual.</td>
</tr>
</tbody>
</table>
Key Terms

**Inferential Statistics:**
Using sample data, together with probability and analysis, in order to make conclusions (i.e. inferences) about a population.

**Research Question:**
The question about the population or long-run process that researchers intend to answer by using data and statistical procedures.

**Representative Sampling Method:**
A sampling method that systematically produces sample statistics that target the value of the population parameter, neither systematically over- or underestimating the true value of the population parameter.

**Biased Sampling Method:**
A sampling method that systematically over- or underestimates the true value of the population parameter.

**Random Sample and Random Sampling Method:**
A method of collecting a sample (called a simple random sample) from a population in such a way that every possible element of the population has an equal chance of being selected. Random sampling is a non-biased sampling method and is a best-practice method for collecting a representative sample.

**Hypothesis Test:**
A statistical procedure in which two opposing hypotheses are proposed, and then probability and observable evidence (i.e. statistics) are used to conclude which hypothesis is plausible.

**Null Hypothesis, \( H_0 \):**
This is the assumption that we make about the value of the population parameter. This assumed value is then used to build a distribution of sample statistics. This hypothesis always contains an equal sign, and the distribution of sample statistics will be centered around this value.

**Alternative Hypothesis, \( H_A \):**
This is a conjecture that is the “opposite” of the null hypothesis. This hypothesis never contains an equal sign. It ONLY contains either < or > or ≠. The alternative hypothesis often represents the research conjecture.

**One-sided Test:**
If the alternative hypothesis contains either < or > then the test is one-sided.

**Two-sided Test:**
If the alternative hypothesis contains ≠ then the test is two-sided.
Inferential Statistics

The research question is the question about the population that researchers intend to answer by using data and statistical procedures. In this chapter, the only research questions that we focus on are those that attempt to understand the value of an unknown population mean, $\mu$, or those that attempt to understand the value of an unknown population proportion, $\pi$. Some example research questions that could be posed related to $\mu$ or $\pi$ include the following:

- What is the average (mean) height of all adult males in the U.S.?
- What is the average (mean) number of credits taken by all students at DCC this semester?
- What is the average (mean) number of minutes an adult female can hold her breath?
- What is the percentage of votes that Candidate Johnson will get in the election?
- What percent of American households own at least one gun?
- What percent of children in America are homeschooled full-time?

In order to create answers to research questions about the population, researchers collect a sample from the population, calculate statistics from the collected sample, and then use those observed statistics, together with probability, in order to make conclusions about the population. This process of using sample data, together with probability and analysis, in order to make conclusions (i.e. inferences) about the population is called inferential statistics.

In algebra, geometry, and other previous math courses, you may have experienced “proving” mathematical statements. For example, you may have been asked to prove that the Pythagorean Theorem is true for all right triangles. In statistics, since our observed evidence is a statistic that is calculated from a collected sample, it is important to understand that there is not going to be a conclusion that is stated as if it is “proof” of the value of a population parameter. We are never going to be able to, for example, take a poll of a sample of voters and then make a conclusion like “this poll proves that Candidate Johnson will get 48% of the votes in the election”. This is not the sort of conclusion that is possible with inferential statistics.

With inferential statistics, we can’t have statements of “proof” because there is always variability among sample statistics. The statistic that we observe from our collected sample is only one of many possible statistics that we could have observed if we had collected a different sample from the population. If we collected a new sample from the population, then we would get a different sample statistic. With inferential statistics, we will take the observed sample statistic and use mathematical probability in order to draw a conclusion about the value of the unknown population parameter. It is important to be very careful in how we phrase our conclusions since our conclusions will never be statements of proof about the value of the unknown population parameter. Instead, you may see the final conclusions about the research questions phrased like the following examples:

- Research Question: What is the average (mean) height of all adult males in the U.S.?
  Conclusion: We are 95% confident that the mean height of all adult males in the U.S. is between 5 feet 8 inches, and 5 feet 10 inches.

- Research Question: What is the average (mean) number of credits taken by all students at DCC this semester?
  Conclusion: It is plausible, at a 5% significance level, that the average (mean) number of credits for all DCC students this semester is 10.4 credits.

- Research Question: What is the average (mean) number of minutes an adult female can hold her breath?

78
Conclusion: We have statistically significant evidence, at a 1% significance level, that the average (mean) number of minutes that an adult female can hold her breath is more than 32 seconds.

- Research Question: What is the percentage of votes that Candidate Johnson will get in the election?
  Conclusion: We have statistically significant evidence, with a p-value of 0.002, that Candidate Johnson will get fewer than 50% of the votes in the election.

- Research Question: What percent of American households own at least one gun?
  Conclusion: It is plausible, with a 5% significance level, that more than 40% of all American households own at least one gun.

- Research Question: What percent of children in America are homeschooled full-time?
  Conclusion: We are 95% confident that the percent of all children who are homeschooled full-time is 3% with a margin of error of ±0.8%.

Notice, in the examples above, that there is always an element of probability included within each conclusion. There are words like “p-value” and “significance level” and “confidence”, which will be discussed and explained within this and later chapters.

In the real world you will, unfortunately, not generally find conclusions phrased in this manner in the headline on the news or in the newspaper. But, if you dig into the actual details of the statistical study on which the headline is based, you will find this information included. The last bullet point above, for example, may only be reported on the news as “3% of Children are Homeschooled!” with the fine print stating something like “margin of error ±0.8%”. When the news reports a “fact” like this, it is important for you, as a consumer, to understand that it isn’t fully accurate to state “3% of Children are Homeschooled”. The true percentage of homeschooled children is unknown, and accurately making a statement about that unknown value is more complicated than what is generally stated on the news.
Introduction to Hypothesis Testing

A hypothesis test is a statistical procedure that is used to create an evidenced-based conclusion about a population parameter. Hypothesis testing is the first inferential statistical procedure that we will learn and practice in this text. Later we will also explore the procedure of creating confidence intervals, which is another commonly used inferential statistical procedure.

Recall that our usual goal is to try to understand/estimate/make a conclusion about the true value of the population parameter. The value of the population parameter is usually unknown since we are generally unable to gather data from every single element of the population. For now, we will only be focused on hypothesis tests involving the population parameters $\pi$ and $\mu$. Recall that $\pi$ is the population proportion of successes for a categorical variable and $\mu$ is the average, mean, of the values of a quantitative variable among members of a population. In the case of $\pi$ and $\mu$, a hypothesis test involves the following basic steps:

**STEP 1.** Set up two contradictory hypotheses:
- The null hypothesis, which makes an assumption about the true value of the population parameter, and
- The alternative hypothesis, which is often based on the research conjecture and makes an opposing statement about the true value of the population parameter.

**STEP 2.** Collect sample data and calculate the sample statistic of interest. Attempt to collect a sample that is a representative sample from the population.

**STEP 3.** Create a distribution of sample statistics (simulated or theory-based) that we can expect to see if the value of the population parameter that was assumed in the null hypotheses were actually the true value of the population parameter.

**STEP 4.** Identify whether the observed statistic is “usual/expected” or “unusual/unexpected” under the assumption about the true value of the population parameter. Conclude one of the following about the population parameter:
- If the assumed value of the population parameter was true, and the observed statistic is “unusual/unexpected” under that assumption, then we conclude that the assumed value of the population parameter is not plausible (believable). We have statistically significant evidence against the null hypothesis, and so we conclude that the alternative hypothesis is plausible (believable).
- If the assumed value of the population parameter was true, and the observed statistics is “usual/expected” under that assumption, then we conclude that the assumed value of the population parameter is plausible (believable). We have little to no evidence against the null hypothesis, and so we conclude that the null hypothesis is plausible (believable).

We will go into more details of step 4 in the next section. For now, we will focus primarily on steps 1, 2, and 3. Details of these steps, as well as examples related to these steps, are below.
Step 1: Writing the Null and Alternative Hypotheses

The first step in hypothesis testing is to write the null and alternative hypotheses. You should get in the habit of writing these hypotheses using the proper notation, and also writing these hypotheses in contextual sentences. Following are the basic features to remember about the null and alternative hypotheses when writing them.

**Null Hypothesis, \( H_0 \):**
- This is a statement that the population parameter is equal to a specific value. It is an assumption we are making about the true value of the population parameter.
- This statement always contains an equal sign.
- The notation for the null hypothesis is \( H_0: \pi = \text{value} \) when the research question aims to understand the value of \( \pi \) (i.e., the variable is categorical).
- The notation for the null hypothesis is \( H_0: \mu = \text{value} \) when the research question aims to understand the value of \( \mu \) (i.e., the variable is quantitative).
- The assumed value of the population parameter will be used to build the distribution of sample statistics. The distribution of sample statistics will be centered around this assumed value.
- We always assume the null hypothesis is true at the start of the hypothesis testing process, and then we make a conclusion at the end of the analysis about whether that assumption is plausible or whether there is statistically significant evidence against the null hypothesis assumption.

**Examples:**
\[ H_0: \pi = 0.8 \] “The null hypothesis is as follows: The proportion of all people who like chocolate is 80%”.
\[ H_0: \mu = 13.2 \] “The null hypothesis is as follows: The average (mean) length of all the boards is 13.2 inches”.

**Alternative Hypothesis, \( H_A \):**
- This is a statement that is the “opposite” of the null hypothesis. At the end of the analysis, you will conclude that either the null hypothesis is plausible, or you will conclude that the alternative hypothesis is plausible.
- This hypothesis never contains an equal sign. It ONLY contains either < or > or \( \neq \).
- This hypothesis always has exactly the same value that was stated in the null hypothesis. The only difference is the use of < or > or \( \neq \). We select the correct symbol based on the research question.
- The notation for alternative hypothesis is \( H_A: \pi < \text{value or } H_A: \pi > \text{value or } H_A: \pi \neq \text{value} \) when the research question aims to understand the value of \( \pi \) (i.e., the variable is categorical).
- The notation for alternative hypothesis is \( H_A: \mu < \text{value or } H_A: \mu > \text{value or } H_A: \mu \neq \text{value} \) when the research question aims to understand the value of \( \mu \) (i.e., the variable is quantitative).

**Examples:**
\[ H_A: \pi < 0.8 \] “The alternative hypothesis is as follows: The proportion of all people who like chocolate is less than 80%”.
\[ H_A: \mu > 13.2 \] “The alternative hypothesis is as follows: The average (mean) length of all the boards is more than 13.2 inches”.
\[ H_A: \pi \neq 0.032 \] “The alternative hypothesis is as follows: The proportion of all people who experienced symptoms is not 3.2%”.

A **one-sided test** is one in which the alternative hypothesis contains either < or >. A **two-sided test** is one in which the alternative hypothesis contains \( \neq \). We will see in the next section that a one-sided test will require that the sample statistics on one side of the distribution of sample statistics are “counted” as evidence against the null hypothesis, whereas a two-sided test will require that the sample statistics on two sides of the distribution of sample statistics are “counted” as evidence against the null hypothesis.
For each research question below, identify the null hypothesis and the alternative hypothesis. Be sure to write both hypotheses in symbols using correct notation, and also in words using complete sentences in context. Also identify whether the test is one-sided or two-sided.

(a) Does Mr. Young have more than 80% support from registered voters?
(b) Is the average length of all the wires from the production line less than 4.2 inches?
(c) Is the new drug still causing the headache side-effect in 2% of patients?
(d) Is the average income for all adults in the state $74,834?

Solution 2.1.1

(a) The hypotheses are as follows:
\[ H_0: \pi = 0.8 \quad 80\% \text{ of all voters will vote for Mr. Young} \]
\[ H_A: \pi > 0.8 \quad \text{More than 80\% of all voters will vote for Mr. Young} \]
This is a one-sided test since the alternative hypothesis uses >.

(b) The hypotheses are as follows:
\[ H_0: \mu = 4.2 \quad \text{The average length of all wires from the production line is 4.2 inches.} \]
\[ H_A: \mu < 4.2 \quad \text{The average length of all wires from the production line is less than 4.2 inches.} \]
This is a one-sided test since the alternative hypothesis uses <.

(c) The hypotheses are as follows:
\[ H_0: \pi = 0.02 \quad \text{2\% of all patients experience headache side effects when taking this drug.} \]
\[ H_A: \pi \neq 0.02 \quad \text{The proportion of all patients with headache side effects when taking this drug is not 2\%.} \]
This is a two-sided test since the alternative hypothesis uses ≠.

(d) The hypotheses are as follows:
\[ H_0: \mu = 74,834 \quad \text{The average annual income for all adults in the state is$74,834.} \]
\[ H_A: \mu \neq 74,834 \quad \text{The average annual income for all adults in the state is not$74,834.} \]
This is a two-sided test since the alternative hypothesis uses ≠.
**Step 2: Collect a Representative Sample**

All of the inferential statistical procedures that we will discuss in this book can only be trusted if the statistics that are used in the process can be trusted to be accurate approximations of the population parameter. We can only trust the observed sample statistic if it came from a sample that was collected by a trusted source, and if the sample was collected in a manner that produces a sample that “looks like” the larger population. The process of sampling will be discussed in more detail later in Chapter 2, but for now we will introduce a few important terms.

A **representative sampling method** is one that systematically produces sample statistics that target the value of the population parameter, neither systematically over- or underestimating the true value of the population parameter. A **biased sampling method** is one that systematically over- or underestimates the true value of the population parameter. A **random sampling method** is a method of collecting a sample (called a **simple random sample**) from a population in such a way that every possible element of the population has an equal chance of being selected. Random sampling is a non-biased sampling method and is a best-practice sampling method for collecting a representative sample. It is important, in all upcoming problems that we consider, for you to note whether the sample that was collected is a random sample, and is therefore a representative sample from the population. If not, then we may not be able to generalize the results of the statistical analysis to the population.

**Step 3: Assume the Null Hypothesis is True and Create a Distribution of Sample Statistics**

Recall that the value of the population parameter is unknown, and the goal of the hypothesis testing procedure is to identify which of the two hypotheses, which make statements about the population parameter, will ultimately be plausible based on the observed evidence. The observed evidence is the sample statistic that was calculated from the collected sample. In order to use the observed sample statistic to make a decision about which hypothesis is plausible, we will assume that the null hypothesis statement is true and then create the distribution of sample statistics that we would expect to see under this assumption. This will allow us, in Step 4, to compare our observed sample statistic to those found in the distribution of sample statistics in order to determine whether our observed sample statistic is usual/expected or unusual/unexpected.

**Step 4: Identify Whether the Observed Statistic is Usual or Unusual, and Make Conclusion**

We will explore this step in more detail in the next section of the text when we explore p-values and discuss the details of how to phrase conclusions. For now, we will simply use the 2 Standard Deviations Rule to determine which sample statistics would be considered usual and which sample statistics would be considered unusual if the null hypothesis were true.
Example 2.1.2

For the scenario below, identify the assumption we should make and then create a distribution of simulated sample statistics based on that assumption. Finally, use the 2 Standard Deviations Rule to describe which sample statistics are usual/expected and which sample statistics are unusual/unexpected under the assumption that the null hypothesis is true.

*Does Mr. Young have more than 80% support from registered voters? A random sample of 114 registered voters is asked whether they support Mr. Young. The hypotheses are as follows:*

\[ H_0: \pi = 0.8 \quad \text{80% of all voters will vote for Mr. Young} \]
\[ H_A: \pi > 0.8 \quad \text{More than 80% of all voters will vote for Mr. Young} \]

**Solution 2.1.2**

The population parameter is the percent of all votes that Mr. Young will get, and the value of this parameter is unknown. The null hypothesis is an assumption about this value. For this hypothesis testing procedure, we will start by assuming that the null hypothesis is true and then we will identify what the sample statistics are expected to look like under that assumption. This means that we will assume it is true that 80% of all voters will vote for Mr. Young. A screenshot from the One proportion inference applet is shown below.

We see from the distribution of simulated sample statistics (distribution of simulated sample p-hats) above that the distribution is centered around the value 0.799, which makes sense because we are assuming that \( \pi = 0.8 \) and the distribution should be centered around that value. We also see that the standard deviation of the distribution is 0.037.

Next, we determine the “cutoffs” between usual and unusual statistics.

\[
0.799 - 0.037 - 0.037 = 0.725 \\
0.799 + 0.037 + 0.037 = 0.873
\]

According to the 2 Standard Deviations Rule, this means that the usual/expected sample statistics (the expected values of \( \hat{p} \)) are those that are between 72.5% and 87.3%. Sample proportions (\( \hat{p} \) values) that are lower than 72.5% or higher than 87.3% are considered unusual/unexpected.

If we collect a sample of size 114 from a population in which 80% of all voters will vote for Mr. Young and the observed sample proportion, \( \hat{p} \), that we calculate from our sample is between 72.5% and 87.3%, then the null hypothesis is plausible. In this case, we would conclude that it is plausible that 80% of all voters will vote for Mr. Young.

But, if we collect a sample of size 114 from a population in which 80% of all voters will vote for Mr. Young and the observed sample proportion, \( \hat{p} \), that we calculate from our sample is higher than 87.3%, then the null hypothesis is not plausible. In this case, we would conclude that we have statistically significant evidence that more than 80% of all voters support Mr. Young.
For the scenario below, identify the assumption we should make and use the 2 Standard Deviations Rule to describe which sample statistics are usual/expected and which sample statistics are unusual/unexpected under the assumption that the null hypothesis is true.

Is the average length of all wires from the production line less than 4.2 inches? A random sample of 78 wires is collected and their lengths are measured. The hypotheses are as follows:

\[ H_0: \mu = 4.2 \quad \text{The average length of all wires from the production line is 4.2 inches.} \]
\[ H_A: \mu < 4.2 \quad \text{The average length of all wires from the production line is less than 4.2 inches.} \]

Solution 2.1.3

The population parameter (the population mean) is the average (mean) length of all wires from the production line, and the value of the parameter is unknown. The null hypothesis is an assumption about this value. For this hypothesis testing procedure, we will start by assuming that the null hypothesis is true and then we will look at the distribution of expected sample means from samples of size 78 if the null hypothesis assumption is true. This means that we will assume it is true that the average (mean) length of all wires from the production line is 4.2 inches.

We see from the distribution of simulated sample statistics (the distribution of simulated sample x-bars) above that the distribution is centered around the value 4.201. This makes sense because we are assuming that \( \mu = 4.2 \) and the distribution should be centered around that value. We also see that the standard deviation is 0.036.

Next, we determine the “cutoffs” between usual and unusual statistics.

\[
4.201 - 0.036 - 0.036 = 4.129 \\
4.201 + 0.036 + 0.036 = 4.273
\]

According to the 2 Standard Deviations Rule, this means that the usual/expected sample statistics (the expected values of \( \bar{x} \)) are those that are between 4.129 inches and 4.273 inches. Sample means (\( \bar{x} \) values) that are lower than 4.129 inches or higher than 4.273 inches are considered unusual/unexpected if the mean length of all wires from the production line is 4.2 inches.

If we collect a sample of 78 wires and the observed sample mean, \( \bar{x} \), is between 4.129 inches and 4.273 inches, then the null hypothesis is plausible. In this case, we would conclude that it is plausible that the average length of all wires from the production line is 4.2 inches.

But, if we collect a sample of 78 wires and the observed sample mean, \( \bar{x} \), is less than 4.129 inches, then the null hypothesis is not plausible. In this case, we would conclude that we have statistically significant evidence that the average length of all wires from the production line is less than 4.2 inches.
Example 2.1.4

A distribution of simulated sample statistics (shown below) was created based on an assumed value of a population parameter. Use the distribution of sample statistics to answer the questions below.

(a) Are we working with categorical data, or are we working with quantitative data in this problem? Explain.

(b) Write the null hypothesis in symbols. Explain what information in the screenshot above was used to help you identify the null hypothesis.

(c) If we switched the applet to display “Proportion of successes” instead of “Number of successes”, then what value would the distribution of sample statistics be centered around? Explain how you know.

(d) Suppose you are now given the additional information that the mean of the distribution, when using the “Proportion of successes” display, is 0.397 and the standard deviation is 0.067. Use that additional information, together with the 2 Standard Deviations Rule, to fill in the blanks in the following sentences.

- It would be unusual to get a sample statistic, $\hat{p}$, that is less than _______ or more than ________.
- The usual/expected sample statistics are those that are between _______ and ________.

(e) Describe the conclusion we should make about the hypotheses based on the observed statistic.

Solution 2.1.4

(a) We are working with categorical data in this problem. We are counting “number of successes” that fall into a certain category.

(b) The null hypothesis is $H_0: \pi = 0.4$. We know this to be true since the value in “Probability of success” in the applet is 0.4, and that is always the place where we enter the assumed value of the population parameter.

(c) If we switched to “Proportion of successes”, then the distribution will be centered at approximately $\pi = 0.4$ since that is the assumed value of the population parameter, and so the sample statistics would be centered around that value.

(d) $0.397 - 0.067 - 0.067 = 0.263$

$0.397 + 0.067 + 0.067 = 0.531$

- It would be unusual to get a sample statistic, $\hat{p}$, that is less than 0.263 or more than 0.531.
- The usual/expected sample statistics are those that are between 0.263 and 0.531.

(e) If we collect a sample of size 50 from a population with $\pi = 0.4$ and the observed sample proportion, $\hat{p}$, is between 26.3% and 53.1%, then the null hypothesis is plausible. In this case, we would conclude that it is plausible that the true value of the population proportion is 40%.

But, if we collect a sample of size 50 from a population with $\pi = 0.4$ and the observed sample statistic, $\hat{p}$, is lower than 26.3% or higher than 53.1%, then the null hypothesis is not plausible. In this case, we would conclude that we have statistically significant evidence that the true value of the population proportion is different than 40% (if the alternative hypothesis included the symbol $\neq$).
A distribution of simulated sample statistics (shown below) was created based on an assumed value of a population parameter. Use the distribution of sample statistics to answer the questions below.

(a) Are we working with categorical data, or are we working with quantitative data in this problem? Explain.

(b) Use the pictured distribution to make your best estimate of the null hypothesis in symbols. Explain what information in the pictured distribution was used to help you identify the null hypothesis.

(c) Use the pictured distribution to make your best estimate of the standard error.

(d) Use your answers from parts (b) and (c) to complete the sentences below.

   *It would be unusual to get a sample statistic, \( \bar{x} \), that is less than ______ or more than _______.
   *The usual/expected sample statistics are those that are between ______ and _______.

(e) Describe the conclusion we should make about the hypotheses based on the observed statistic.

Solution 2.1.5

(a) We are working with quantitative data in this problem. Each square in the distribution represents a sample mean (as labeled on the horizontal axis), which can only be calculated from quantitative data.

(b) It appears that the null hypothesis is \( H_0: \mu = 315 \). This is because the value in the center of the distribution appears to be 315. Since this is a distribution of simulated sample means, the distribution is centered about the population mean, \( \mu \).

(c) Since the distribution is roughly bell-shaped, we can use our understanding of the length of 1 standard deviation on a bell-shaped distribution to approximate the size of the standard error. Sample means 312 and 318 appear to be approximately 2 standard deviations away from the center, which means that the standard deviation of the distribution (i.e., the standard error) is approximately 1.5 \( \left( \frac{318 - 315}{2} = \frac{3}{2} = 1.5 \right) \).

(d) \[ 315 - 1.5 - 1.5 = 312 \quad 315 + 1.5 + 1.5 = 318 \]

   *It would be unusual to get a sample statistic, \( \bar{x} \), that is less than 312 or more than 318.*

   *The usual/expected sample statistics are those that are between 312 and 318.*

(e) If we collect a sample from a population with a mean of 315 and the observed sample mean, \( \bar{x} \), is between 312 and 318, then the null hypothesis is plausible. In this case, we would conclude that it is plausible that the true value of the population mean is 315.

But, if we collect a sample from a population with a mean of 315 and the observed sample mean, \( \bar{x} \), is higher than 318 or lower than 312, then the null hypothesis is not plausible. In this case, we would conclude that we have statistically significant evidence that the true value of the population proportion is different than 315 (if the alternative hypothesis included the symbol \( \neq \)).
Section 2.1 | Introduction to Hypothesis Testing

Exercises

1) In each scenario below, describe the population and the parameter of interest in words and determine whether statistics calculated from samples collected as indicated are likely to overestimate, underestimate, or target the true value of the parameter. Explain your reasoning.
   (a) A voluntary email survey is sent to all high school students in the United States to determine the proportion of teens that have used illegal drugs.
   (b) A daytime television show asks its viewers to indicate their favorite channel to watch through a voluntary call-in survey.
   (c) A survey was sent to all DCC students via email to determine the proportion of students with reliable internet access.
   (d) A survey of a random selection of current DCC students asked students to state the number of hours spent the previous week in either the library or a tutoring center. The surveys are conducted (and responses required) via one of their current instructors.

2) A random sample of 100 people living in the United States is collected.
   (a) Is it possible for the sample to be representative of the population of the United States if the variable of interest is “shoe size”? Explain your reasoning.
   (b) Is it plausible for the sample to be representative of the population of the United States if the variable of interest is “shoe size”? Explain your reasoning.

3) Josh is a foosball player who is working on his trick shots. While practicing, he gets 12 out of 20, or 60%, of his trick shots in the goal. Josh wants to know whether his long-run proportion of making his trick shots is greater than 50%.
   (a) Describe the parameter of interest in this study and assign it the appropriate symbol.
   (b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

4) You forgot which of your dice are weighted and decide to conduct a test to determine whether the twelve-sided die you picked up is one of your weighted dice. You roll it a total of 83 times and record the results. If the die is weighted it should roll a 10, 11, or 12 more frequently than a standard 12-sided die. Out of the 83 rolls, you record a 10, 11, or 12 a total of 31 times.
   (a) Describe the parameter of interest in this study and assign it the appropriate symbol.
   (b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.
   (c) What conclusion would you draw from the distribution of simulated sample statistics shown at right?
Kelley Blue Book is a vehicle valuation and automotive research company that reports market value prices for new and used automobiles. These values are often used by auto dealerships to determine the selling prices of their vehicles. The average Kelley Blue Book value of a used 2018-2020 SUV is $19,560. I would like to determine whether I should ask for a different selling price in my region for my 2019 SUV, so I conduct a random sample of local ads for used 2018-2020 SUVs and record the asking price for 17 of them, obtaining an average asking price of $16,880.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?

It has been claimed that in some sports there might be a competitive advantage for teams wearing blue uniforms. You want to test this with the sport of competitive swimming. You conduct a study using 35 separate races where 3 competitive swimmers are chosen at random and assigned either a blue, red, or green swimsuit for each race. The swimmer wearing a blue swimsuit won in 16 of the 35 races.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?
7) The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all men ages 20-29 to be 5.8 feet (5 feet 9.5 inches) with a standard deviation of 0.35 feet (4.2 inches). You want to determine whether the average height of Dutchess County men ages 20-29 is less than the national average. You collect a random sample of 110 Dutchess County residents between the ages of 20 and 29. Of the 110 sampled, 42 identified as men and measured an average height of 5.75 feet (5 feet 9 inches).

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?

8) The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all women ages 20-29 to be 5.375 feet (5 feet 4.5 inches) with a standard deviation of 0.283 feet (3.4 inches). You want to determine whether the average height of Dutchess County women ages 20-29 is greater than the national average. You survey a random sample of 110 Dutchess County residents ages 20-29. Of the 110 sampled, 53 identified as women and measured an average height of 5.485 feet (5 feet 5.8 inches).

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?
9) A study was conducted by the National Endowment for the Arts to describe music preferences in the Hudson Valley. The 2010 report describes relationships between key demographic characteristics and music preferences. In this report, it was noted that in 2010, 24% of adults preferred the genre of Rock/Heavy Metal. Recently, a random sample of 715 Hudson Valley adults was surveyed with 150 indicating they prefer Rock/Heavy Metal. You want to determine whether the proportion of Hudson Valley residents that prefer Rock/Heavy Metal has changed since the 2010 report.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?

10) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that 2 out of every 3 people prefer their brand of cola. A random sample of 140 people participate in the taste test, and 85 of them choose Pepsi as their preferred cola.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?
11) In 1948 the Chicago Daily Tribune erroneously published a headline due to early polling results that indicated Thomas E. Dewey had defeated Harry S. Truman in that year’s presidential election. In truth, Truman won 49.6% of the popular vote (24,179,347 out of 47,346,569 votes) to Dewey’s 45.1% (21,991,292 out of 47,346,569 votes).

The Gallup’s September 24 report indicated that they had obtained a random sample of 3,250 voters, and 46.5% of sampled voters indicated they would be voting for Dewey, while only 38% planned to vote for Truman.

(a) What is the parameter of interest for Truman’s campaign election team?

(b) Truman’s campaign election team wishes to investigate the research question, is 38% of the voting population planning to select Truman, or will Truman receive more than 38% of the vote? Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What is the parameter of interest for Dewey’s campaign election team?

(d) Dewey’s campaign election team wishes to investigate the research question, is 46.5% of the voting population planning to select Dewey, or will Dewey receive fewer than 46.5% of the vote? Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

12) The United States Health and Human Services Poverty Guidelines for 2017 are indicated below:

<table>
<thead>
<tr>
<th>Family/Household Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty Threshold</td>
<td>$12,060</td>
<td>$16,240</td>
<td>$20,420</td>
<td>$24,600</td>
<td>$28,780</td>
<td>$32,960</td>
<td>$37,140</td>
</tr>
</tbody>
</table>

In March 2017 a PBS report stated that approximately 8.3% of citizens live below the poverty line in the United States of America. A study is conducted to determine whether the proportion of New York State residents living below the poverty line is greater than 8.3%. In a random sample of 10,000 New York State residents, it is determined that 1,061 (10.61%) were living below the poverty line in 2017.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.

(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.

(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?
13) A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest was for workers who live in New York City, who traveled an average of 35.9 minutes each day (or a total of 89.4 hours travel time per year) with a standard deviation of 12.5 minutes. You live relatively close to the city and know that many of your neighbors work there, so you decide to investigate whether people in your neighborhood have a longer average commute time than the average worker who lives in New York City. You survey a total of 27 commuters in your neighborhood and find an average commute time of 43.7 minutes.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.
(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.
(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?

14) A 2017 survey conducted by the National Science Foundation claimed that 15% of those currently employed in a STEM (Science, Technology, Engineering, and Mathematics) field in the United States are under the age of 30. You conduct a similar survey to determine whether a smaller proportion of STEM field workers in your state are under the age of 30. After collecting a random sample of 500 STEM field workers in your state, you find that 13.2% of those surveyed are under the age of 30.

(a) Describe the parameter of interest in this study and assign it the appropriate symbol.
(b) Determine the null hypothesis and the alternative hypothesis for this scenario. Write each hypothesis in both words and symbols.
(c) What conclusion would you draw based on the distribution of simulated sample statistics shown at right?
Section 2.2 | Introduction to P-value and Conclusions in a Hypothesis Test

Key Terms

**Plausible:**
An event that is plausible is reasonably likely to occur. An event that is plausible is believable.

**Statistically significant:**
Given a set of assumptions, an event is statistically significant if it is unlikely to occur by random chance under that set of assumptions.

**P-value:**
The p-value is the probability of getting a sample statistic at least as extreme as the one observed, in the direction of the alternative hypothesis, when the null hypothesis is true.

- **NOTE:** Getting a “small” p-value means that, if the null hypothesis were true, then there would be a “small” probability of getting the sample statistic like the one that was observed in the sample. Since there is a “small” probability of getting what was observed when the null hypothesis is true, we would conclude that the null hypothesis is not plausible and should be rejected. In this case, we conclude that the alternative hypothesis is plausible (and the null hypothesis is not plausible). We have statistically significant evidence to support the alternative hypothesis.

**Significance Level:**
The significance level, denoted with the symbol \( \alpha \) (alpha), is the “cut-off” probability that is used to determine whether we reject the null hypothesis (and conclude that the alternative hypothesis is plausible) or conclude that the null hypothesis is plausible.

- If the p-value is smaller than \( \alpha \), then we reject the null hypothesis and conclude that the alternative hypothesis is plausible. We have statistically significant evidence to support the alternative hypothesis.
- If the p-value is bigger than \( \alpha \), then we conclude that the null hypothesis is plausible (i.e., we don’t have sufficient evidence to reject the null hypothesis).
- The significance level is the probability of rejecting the null hypothesis when it is actually true.
Writing an Accurate Conclusion in a Hypothesis Test

“Plausible” vs. “Possible”
It is important to remember that “plausible” and “possible” are not the same thing. We can say “anything is possible”, even if there is only an extremely small chance of that event ever happening. But, if an event only has a small chance of occurring, then we certainly shouldn’t conclude that it is a plausible event.

• In statistics, an event’s likelihood is considered a “small chance” (and therefore not considered a plausible event) when its probability of happening is smaller than the significance level for that statistical study.

The word “plausible” is reserved for circumstances or events that are more than “possible”, but instead are reasonably likely to occur.

• Possible: “could possibly happen even if just by chance,” “might happen,” “can exist”

• Plausible: “believable,” “seems like it could be true,” “credible,” “has appearance of truth”

“Plausible” vs. “Definitely True”
It is important to remember that concluding that an event is plausible is not the same as saying that the event will definitely occur.

Suppose we make an assumption about the value of a population parameter, and then we create a distribution to display the sample statistics we would expect to see if that assumption were true. We can then calculate a sample statistic from a randomly collected sample and compare our observed sample statistic to those found in the distribution of sample statistics in order to make a conclusion about the plausibility of our assumption regarding the population parameter.

• If our observed sample statistic is usual/expected in the distribution of sample statistics, then we say that our assumption about the value of the population parameter is plausible.

• If our observed sample statistic is usual/expected in the distribution of sample statistics, then we do NOT say that our assumption about the value of the population parameter is definitely true!

For example, suppose we make an assumption that 64% of all people in the population prefer Product A when given a choice of products. This means that we are assuming that the true value of the population parameter is equal to \( \pi = 0.64 \). We could then create a distribution of sample statistics in order to view the sample statistics that we would expect to see if it were true that the population parameter was equal to \( \pi = 0.64 \). Finally, we could collect a random sample, and calculate an observed sample statistic and compare it to the distribution of sample statistics in order to determine whether the observed sample statistic is usual or unusual. If the observed sample statistic is usual/expected in the distribution, then

• we conclude that “it is PLAUSIBLE that 64% of people prefer Product A”.

• we can NOT conclude that it is definitely true that 64% of people prefer Product A.
Note that the fact that our statistic is usual/expected does NOT prove that \( \pi = 0.64 \). The observed statistic is only close enough to the value that we assumed for the population parameter to allow us to conclude that the value that we assumed for the population parameter (\( \pi = 0.64 \)) is plausible. Other values of the population parameter are also plausible (we’ll learn more about this later in the text when we discuss confidence intervals). We can only conclude that this particular value of the parameter is plausible. The statistic that we observed tells us that it is plausible that the true value of the population parameter is \( \pi = 0.64 \).

**Statistically significant:**
Given a set of assumptions, an event is statistically significant if it is unlikely to occur by random chance under that set of assumptions. Remember that the distribution of sample statistics shows us what statistics we should expect under a given assumption about the value of the population parameter. Once the distribution of sample statistics is created, we can check to see if our observed statistic is

1) a usual/expected sort of statistic to get under that assumption about the population parameter.

or

2) an unusually high or unusually low (i.e. statistically significant) statistic to get under that assumption about the population parameter.

**Introduction to P-value and Significance Level**

Remember that the goal of the hypothesis testing process is to make a conclusion about the unknown value of the population parameter. To complete the hypothesis testing process, we create null and alternative hypotheses, collect a representative sample, create a distribution of sample statistics based on the assumption that the null hypothesis is true, use that distribution to determine whether our observed sample statistic is usual or unusual, and make a conclusion about which hypothesis is plausible.

In the previous section, we used the 2 Standard Deviations Rule to decide which statistics were usual and unusual. The more common practice is to use a p-value together with a significance level in order to make a decision about which hypothesis is plausible.

The probability of getting a sample statistic at least as extreme as the one observed (in the direction of the alternative hypothesis), assuming that the null hypothesis is true, is called the **p-value**.

The **significance level**, denoted by the symbol \( \alpha \) (alpha), is the “cut-off” probability for determining whether we reject the null hypothesis (conclude that the alternative hypothesis is plausible) or fail to reject the null hypothesis (conclude that the null hypothesis is plausible). In other words, the significance level, \( \alpha \), determines which p-values are considered “small” enough to conclude that the observed statistic is “unusual.”

- If the p-value is smaller than \( \alpha \), then we reject the null hypothesis and conclude that the alternative hypothesis is plausible. We have statistically significant evidence to support the alternative hypothesis. **p-value is small** (smaller than \( \alpha \)) \( \leftrightarrow \) observed statistic is unusual \( \leftrightarrow \) reject \( H_0 \) \( \leftrightarrow \) support \( H_A \)
• If the p-value is bigger than $\alpha$, then we fail to reject the null hypothesis and we conclude that the null hypothesis is plausible. We don’t have sufficient evidence to support the alternative hypothesis.

\[
p\text{-value is big (bigger than } \alpha) \iff \text{observed statistic is usual} \iff H_0 \text{ is plausible}
\]

### Example 2.2.1

A production line is supposed to produce boxes of cereal with an average of 20 ounces of cereal in the boxes. The line supervisor wants to know whether the boxes are coming out with an average of 20 ounces like they are supposed to. The hypotheses are as follows:

- $H_0$: $\mu = 20$ All of the boxes have an average (mean) of 20 ounces of cereal.
- $H_A$: $\mu \neq 20$ The average (mean) amount of cereal in all boxes is not 20 ounces.

The supervisor collects a random sample of 65 boxes from the production line, and uses the observed sample statistic, $\bar{x} = 19.8$ ounces, to find a p-value of 0.24.

(a) Write a sentence to interpret the meaning of the p-value in context.

(b) Which hypothesis is plausible if using a significance level of $\alpha = 0.05$?

#### Solution 2.2.1

(a) If the null hypothesis is true, then the probability of getting a sample statistic at least as extreme as the observed statistic is the p-value. In context, we would say this as follows:

> If it is true that all the boxes have an average (mean) of 20 ounces of cereal, then the probability of getting a random sample of 65 boxes with an average of at least 19.8 ounces is 24%.

(b) The p-value is 0.24 and the significance level is $\alpha = 0.05$. Since the p-value is bigger than the significance level, we conclude that the null hypothesis is plausible. Based on this sample of 65 boxes, it is plausible that all the boxes have an average (mean) of 20 ounces of cereal.

### Example 2.2.2

An invasive plant has shown up in a certain county. A grant funding agency will award funds for the county to combat the invasive species only if more than 10% of the trees in the area are impacted. The hypotheses are as follows:

- $H_0$: $\pi = 0.1$ 10% of all trees in the county are impacted.
- $H_A$: $\pi > 0.1$ More than 10% of all trees in the county are impacted by the invasive species.

A random sample of 108 trees in the county is examined by researchers, and the sample statistic, $\hat{p} = 0.1759$, is used to find a p-value of 0.007.

(a) Write a sentence to interpret the meaning of the p-value in context.

(b) Which hypothesis is plausible if using a significance level of $\alpha = 0.01$?

#### Solution 2.2.2

(a) If the null hypothesis is true, then the probability of getting a sample statistic at least as extreme as the observed statistic is the p-value. In context, we would say this as follows:

> If it is true that 10% of all trees in the county are impacted by the invasive species, then the probability of finding 17.59% or more trees impacted when sampling 108 trees is 0.7%.

(b) The p-value is 0.007 and the significance level is $\alpha = 0.01$. Since the p-value is smaller than the significance level, we conclude that the alternative hypothesis is plausible. Our observed statistic gives statistically significant evidence that more than 10% of trees in the county are impacted by the invasive species.
According to one polling agency, which conducted a poll of random registered voters throughout the U.S. on 9/8/2020, the percent of all voters in the U.S. who approved of the president on that date was 47%.

- Li Wei believes that the registered voters in his area of the country have a lower presidential approval rate, so he conducts a random sample of registered voters in his area. After conducting his random sample, his sample statistic produces a p-value of 0.08.

- Olivia believes that the registered voters in her area of the country have a lower presidential approval rate than 47%, and so she conducts a random sample of registered voters in her area. After conducting her random sample, her sample statistic produces a p-value of 0.02.

(a) Get in groups of 3 students. As a group, answer the following questions. Make sure that everyone in the group agrees with each answer before moving to the next question.

(b) Complete the following questions for Li Wei’s sample.
   i. Write the null and alternative hypotheses for Li Wei using the correct symbols, and also write the hypotheses using complete sentences in context.
   ii. Write a sentence to interpret the meaning of the p-value in context.
   iii. Suppose Li Wei is using a significance level of 0.05. Identify the symbol that would be used to represent the 0.05.
   iv. Using the significance level of 0.05, identify which of the following conclusions is appropriate: “Li Wei should conclude that \( H_A \) is plausible” or “Li Wei should conclude that \( H_0 \) is plausible”. Explain how you know which conclusion to select.
   v. Write a complete sentence to give the conclusion that Li Wei should make in context.

(c) Complete the following questions for Olivia’s sample.
   i. Write the null and alternative hypotheses for Olivia using the correct symbols, and also write the hypotheses using complete sentences in context.
   ii. Write a sentence to interpret the meaning of the p-value in context.
   iii. Suppose Olivia is using a significance level of 0.05. Identify the symbol that would be used to represent the 0.05.
   iv. Using the significance level of 0.05, identify which of the following conclusions is appropriate: “Olivia should conclude that \( H_A \) is plausible” or “Olivia should conclude that \( H_0 \) is plausible”. Explain how you know which conclusion to select.
   v. Write a complete sentence to give the conclusion that Olivia should make in context.

(d) For each statement below, identify whether the statement is valid and explain why or why not.
   i. It is possible that the presidential approval rate in Li Wei’s area of the country is 47%.
   ii. It is plausible that the presidential approval rate in Li Wei’s area of the country is 47%.
   iii. We have statistically significant evidence that the presidential approval rate in Li Wei’s area of the country is 47%.
   iv. The presidential approval rate in Li Wei’s area of the country is 47%.
   v. It is possible that the presidential approval rate in Li Wei’s area of the country is less than 47%.
   vi. It is plausible that the presidential approval rate in Li Wei’s area of the country is less than 47%.
   vii. We have statistically significant evidence that the presidential approval rate in Li Wei’s area of the country is less than 47%.
   viii. The presidential approval rate in Li Wei’s area of the country is less than 47%.

(e) For each statement below, identify whether the statement is valid and explain why or why not.
   i. It is possible that the presidential approval rate in Olivia’s area of the country is 47%.
   ii. It is plausible that the presidential approval rate in Olivia’s area of the country is 47%.
   iii. We have statistically significant evidence that the presidential approval rate in Olivia’s area of the country is 47%.
   iv. The presidential approval rate in Olivia’s area of the country is 47%.
   v. It is possible that the presidential approval rate in Olivia’s area of the country is less than 47%.
   vi. It is plausible that the presidential approval rate in Olivia’s area of the country is less than 47%.
   vii. We have statistically significant evidence that the presidential approval rate in Olivia’s area of the country is less than 47%.
   viii. The presidential approval rate in Olivia’s area of the country is less than 47%.
Smaller P-Value Indicates Stronger Evidence Against the Null Hypothesis

It is extremely common for students to mix up the hypothesis that is plausible based on the p-value. Students often mistakenly think that a bigger p-value implies stronger evidence against the null hypothesis, but this is backwards!

**Facts about the p-value:**

- The smaller the p-value, the stronger the evidence against the null hypothesis.

- The p-value is the probability of getting a sample statistic at least as extreme as the one observed if the null hypothesis is true. A smaller probability indicates that the observed statistic is unlikely if the null hypothesis is true. Thus, a smaller p-value suggests that the null hypothesis assumption is not plausible.

- A smaller p-value indicates that the observed statistic is further out in the tail of the null distribution, which makes the observed statistic more unusual if the null hypothesis is true. A smaller p-value indicates that the null hypothesis is not plausible.

**How to Calculate the P-value Given a Distribution of Simulated Sample Statistics**

The p-value is the probability of getting a sample statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. A p-value smaller than $\alpha$ tells us that the alternative hypothesis is plausible, and a p-value bigger than $\alpha$ tells us that the null hypothesis is plausible. We will now learn how to find a p-value with the help of an applet.

- To find the p-value with a given distribution of sample statistics, we count the number of sample statistics that are equal to the observed sample statistic or more extreme, and then divide by the total number of sample statistics in the distribution. This gives the proportion of sample statistics that are at least as extreme as the observed statistic.

- “More extreme” means that we count statistics in the direction of the alternative hypothesis.

  - If the alternative hypothesis is $H_A$: $parameter > value$, then we will count sample statistics at or above the observed sample statistic.

  - If the alternative hypothesis is $H_A$: $parameter < value$, then we will count sample statistics at or below the observed sample statistic.

  - If the alternative hypothesis is $H_A$: $parameter \neq value$, then we will find the distance that the sample statistic is from the center (mean), and then count sample statistics like the observed sample statistic or further into the tail in both directions.
Example 2.2.3

Suppose the hypotheses in a statistical study are as follows:

\( H_0: \pi = 0.4 \)  
40% of all employees at the company walk to work.

\( H_A: \pi < 0.4 \)  
Fewer than 40% of all employees at the company walk to work.

Suppose we collect a random sample of all employees at the company and calculate a sample statistic of 
\( \hat{p} = \frac{12}{50} = 0.24 = 24\% \). We then use the One proportion inference applet to create a distribution of 100 simulated sample statistics, and the results are pictured below.

a) If we switched to “Proportion of successes” for the graph, at what value would the distribution be centered? What value is this distribution centered at currently (using the “Number of successes” option)? How do we know the value of the center without even selecting the “Summary Stats” button?

b) Find the p-value. Explain how to use the given distribution to find the p-value.

c) Write a sentence to interpret the p-value in context.

d) What conclusion do we make about the hypotheses if we are using a significance level of \( \alpha = 0.05 \)?

Solution 2.2.3

a) The distribution would be centered at approximately \( \pi = 0.4 \) since that is the value assumed in the null hypothesis. Currently, when using the “Number of successes” option, the distribution is centered at approximately 20 since the sample size is 50, and we expect the distribution to be centered at approximately 40% of 50 which is \( 0.4 \times 50 = 20 \).

b) The p-value is found by counting all the simulated sample statistics (dots) that are at or below the observed statistic. There are 100 total dots since the applet produced 100 simulated sample repetitions. The observed sample statistic is 12 successes out of 50. The alternative hypothesis is that \( \pi < 0.4 \), and because the alternative hypothesis uses the < symbol, we will count the dots that are like the observed statistic or lower. So, we count the dots that are 12 or lower (circled in the screenshot below). Thus, the p-value is \( 4 \text{ dots}/(100 \text{ total dots}) = 0.04 = 4\% \).

c) If it is true that 40% of all employees at the company walk to work, then the probability of getting a sample of 50 employees with 12 or fewer who walk to work is 4%.

d) Since the p-value is smaller than the significance level (0.04 is smaller than 0.05), we conclude that the alternative hypothesis is plausible. We conclude that it is plausible that fewer than 40% of all employees at the company walk to work.
Example 2.2.4

Suppose the hypotheses in a statistical study are as follows:

\( H_0: \mu = 315 \) The population of all students in the school district spent an average (mean) of 315 minutes using screens during the weekend.

\( H_A: \mu \neq 315 \) The average (mean) amount of time the population of all students in the school district spent on screens this weekend was not 315 minutes.

Suppose we collect a random sample of 45 students in the school district and calculate the mean amount of screen time to be \( \bar{x} = 318 \) minutes. We then use the One Variable with Sampling applet to create 100 simulated sample statistics, as pictured below.

![Histogram of sample means with 100 simulated samples]

a) At what value is this distribution centered? How do we know the value of the center without even selecting the “Summary stats” button?

b) Find the p-value. Explain how to use the given distribution to find the p-value.

c) Write a sentence to interpret the p-value in context.

d) What conclusion do we make about the hypotheses if we are using a significance level of \( \alpha = 0.01 \)?

Solution 2.2.4

a) The distribution would be centered at approximately \( \mu = 315 \) since that is the value assumed in the null hypothesis.

b) The p-value is found by counting all the sample statistics (the squares) that are at least as extreme as the observed statistic. The observed sample statistic is \( \bar{x} = 318 \). The alternative hypothesis is that \( \mu \neq 315 \), and because the alternative hypothesis uses the \( \neq \) symbol, we will count the dots that are like our observed statistics or further away from the center. On the right side, 318 is 3 units above the center of the distribution. On the left side, the number 312 is 3 units below the center of the distribution. We will count squares that are 318 and higher, and also count squares that are 312 or lower. There are 6 squares that are 318 or higher, and there are 7 squares that are 312 or lower (circled in the screenshot below). The p-value is \( \frac{6 \text{ squares} + 7 \text{ squares}}{100 \text{ total squares}} = \frac{13 \text{ squares}}{100 \text{ total}} = 0.13 = 13\% \).

![Screenshot showing 6 squares above and 7 squares below the center]

c) If it is true that the population of all students in the school district spent an average (mean) of 315 minutes using screens during the weekend, then the probability of getting a sample of 45 students who have an average screen time of 318 minutes or more extreme (in both directions) is 13%.

d) Since the p-value is bigger than the significance level (0.13 is bigger than 0.01), we conclude that the null hypothesis is plausible. We conclude that it is plausible that the students in the school district spent an average (mean) of 315 minutes using screens during the weekend.
A certain candy company claims that 18% of all the candies they manufacture are blue. Imani doesn’t like the blue candies, and she thinks that the bags that she buys have more than 18% blue candies. She takes a random sample of 50 candies, and she finds that 16 of the candies in her sample are blue. Imani then uses the One proportion inference applet to create a distribution of 100 simulated sample statistics, which is pictured below.

Work together to answer each of the questions below. Be ready to share your answers with the class.

(a) Write the null and alternative hypotheses in both words and symbols.
(b) At approximately what value is the simulated distribution of sample statistics centered? If we switched the applet to display “Proportion of successes,” then what would be the value of the center? Explain how you know.
(c) Find the p-value using the distribution pictured above. Explain your process.
(d) Write a sentence to interpret the p-value.
(e) What conclusion should Imani make about the hypotheses if she is using a significance level of \( \alpha = 0.05 \)?

(f) For each statement below, identify whether the statement is valid and explain why or why not.
   i. It is possible that 18% of all the candies made by the company are blue.
   ii. It is plausible that 18% of all the candies made by the company are blue.
   iii. We have statistically significant evidence that 18% of all the candies made by the company are blue.
   iv. 18% of all the candies made by the company are blue.
   v. It is possible that more than 18% of all the candies made by the company are blue.
   vi. It is plausible that more than 18% of all the candies made by the company are blue.
   vii. We have statistically significant evidence that more than 18% of all the candies made by the company are blue.
   viii. More than 18% of all the candies made by the company are blue.

How to Calculate the p-value using an Applet

Instead of counting dots or squares by hand, the Rossman Chance applets can calculate the p-value for us.

- To find the p-value using an applet, we need the applet to count simulated sample statistics that are at least as extreme as the observed statistic.

- For hypothesis tests involving \( \pi \) (i.e., the variable is categorical), in the One proportion inference applet:
  - Select the correct “As extreme as” symbol that matches the direction of the alternative hypothesis symbol.
    (either \( \geq \) or \( \leq \) or select the “Two-sided” box for \( \neq \)).
  - Enter the value of the observed statistic in the “Count” box.
  - Click Count.
• For hypothesis tests involving $\mu$ (i.e., the variable is quantitative), in the One Variable with Sampling applet:

  - Select the correct “As extreme as” symbol that matches the direction of the alternative hypothesis symbol (either $\geq$ or $\leq$ or select “Beyond” for $\neq$).
  - Enter the value of the observed statistic in the “Count Samples” box.
  - Click Count.

• The applet will display a red vertical line, and also color in red the simulated sample statistics that are being counted. The applet will also display the calculated p-value in red.

The observed statistic is $\hat{p} = 0.72$.
The alternative hypothesis is $H_A: \pi > 0.6$.

There is a red vertical line with the p-value displayed at the top of the line. In this case, the p-value is $(8 \text{ dots})/(100 \text{ total}) = 0.08$.
The p-value calculation and value are also displayed in red underneath the “As extreme as” area.

The observed statistic is $\bar{x} = 99$
The alternative hypothesis is $H_A: \mu < 100$.

There is a red vertical line and the squares that are 99 or lower are highlighted in red to indicate that they are being counted.
The p-value calculation and value are displayed in red underneath the “Count Samples” area. The p-value is 0.07.
Example 2.2.5

A person wants to know whether his six-sided die is fair. To test the die, he rolls it 60 times and records the number of times that the die rolls the number six. It turns out that, in the 60 rolls, the die rolled a six a total of 5 times. The hypotheses for this scenario are as follows:

\[ H_0: \pi = \frac{1}{6} \approx 0.167 \quad \text{In the long run, the die rolls the number six} \frac{1}{6} \text{ of the time (the die is fair).} \]

\[ H_A: \pi \neq \frac{1}{6} \quad \text{In the long run, the die rolls the number six at a rate different than} \frac{1}{6} \text{ of the time (the die is not fair).} \]

(a) Use the One proportion inference applet to create a distribution of 1000 simulated sample statistics, and also use the applet to calculate the p-value. Include a screenshot from the applet.

(b) Interpret the meaning of the p-value in context.

(c) Write a conclusion about the hypotheses in context if the significance level is 0.05.

Solution 2.2.5

(a) The sample statistic is \( \hat{p} = \frac{5}{60} \approx 0.083 = 8.3\% \). A distribution of 1000 simulated sample statistics from the applet is shown below:

The parameter value assumed in the null hypothesis is entered in the “Probability of success” box, and the distribution is centered at this value.

The die was rolled 60 times in each sample.

There are 1000 sample statistics simulated.

(b) If it is true that the die is fair, then the probability of rolling the number six 5 times or less, when rolling the die 60 times, is 10.2%.

(c) Since the p-value is bigger than the significance level, we conclude that the null hypothesis is plausible. This means that it is plausible that the die is fair.

Note: in the example above, if we switched to “Number of successes”, then we could put the number 5 in the “Count” box in order to achieve the same results.
Example 2.2.6

Candace makes a conjecture about the average (mean) GPA of all DCC students. She then collects a random sample of 120 DCC students and calculates the average (mean) GPA of the students in her sample. Finally, she uses the One Variable with Sampling applet to create the distribution of simulated sample statistics shown below.

(a) Use the information shown in the screenshot to write the null and alternative hypotheses in both words and symbols.
(b) Use the information in the screenshot to identify the value of the observed sample statistic, including the correct symbol.
(c) Use the information in the screenshot to identify the p-value.
(d) Interpret the meaning of the p-value in context.
(e) Write a conclusion about the hypotheses in context if the significance level is 0.05.

Solution 2.2.6

(a) The distribution is centered at a value of 2.9, so that must be the assumed value of the population mean. The p-value is calculated by counting simulated statistics that are like the observed statistic or higher, so that tells us that the alternative hypothesis includes the symbol \( > \). The hypotheses are as follows:

\[ H_0 : \mu = 2.9 \quad \text{The average GPA of all DCC students is 2.9.} \]
\[ H_0 : \mu > 2.9 \quad \text{The average GPA of all DCC students is more than 2.9.} \]

(b) The observed sample statistic is a sample mean with value \( \bar{x} = 3.02 \).

(c) The p-value is the proportion of sample statistics that are at or above the observed sample statistic. The applet counts 4 simulated sample statistics out of a total of 100. The p-value is \( \frac{4}{100} = 0.04 = 4\% \).

(d) If it is true that the average (mean) GPA of all DCC students is 2.9, then the probability of getting a sample of 120 DCC students who have an average GPA of 3.02 or higher is 4%.

(e) Since the p-value is smaller than the significance level, we conclude that the alternative hypothesis is plausible. There is statistically significant evidence that the average (mean) GPA of all DCC students is higher than 2.9.
Section 2.2 | Introduction to P-value and Conclusions in a Hypothesis Test

Exercises

1) Compare and contrast the terms possible and plausible.

2) Give an example of an event that is possible, but not plausible.

3) Given the distribution of simulated statistics below, determine the p-value for each set of hypotheses and sample statistic. Explain the process you used to obtain the p-value.

(a) \( H_0: \pi = 0.4 \)
   \( H_A: \pi > 0.4 \)
   \( \hat{p} = 0.55 \)

(b) \( H_0: \pi = 0.4 \)
   \( H_A: \pi < 0.4 \)
   \( \hat{p} = 0.2 \)

(c) \( H_0: \pi = 0.4 \)
   \( H_A: \pi \neq 0.4 \)
   \( \hat{p} = 0.6 \)

4) Complete the following statements by choosing the correct term in parentheses.

(a) A p-value smaller than \( \alpha \) means we should ___________ (accept/reject) the ___________ (null/alternative) hypothesis.

(b) The significance level is the probability of rejecting the ___________ (null/alternative) hypothesis when it is true.

(c) If the p-value is larger than the significance level, then we can say that the assumption about the parameter is ________________ (plausible/not plausible).

(d) If the sample statistic is unusual/unexpected, then we can say that the assumption about the population parameter is ________________ (plausible/not plausible)
5) It has been claimed that in competitive sports there might be some advantage for teams wearing blue uniforms. You want to test this with the sport of competitive swimming. You conduct a study using 35 separate races where 3 competitive swimmers are chosen at random and assigned either a blue, red, or green swimsuit for each race. The swimmer wearing a blue swimsuit won in 16 of the 35 races.

(a) Describe the parameter of interest in this study and assign the appropriate symbol.
(b) Write the null and alternative hypotheses for this scenario in both words and symbols.
(c) Determine the values and inequality symbol that should be entered in the One proportion inference applet (shown to the right) to create the distribution of simulated sample statistics shown below.

(d) What conclusion should you draw using a significance level of 1%?
(e) What conclusion should you draw using a significance level of 5%?
(f) What conclusion should you draw using a significance level of 10%?

(g) Complete the following statement:

Assuming swimsuit colors provide no competitive advantage when racing, the probability of the swimmers wearing blue winning _______% or more of a set of 35 races is _______%.
Kelley Blue Book is a vehicle valuation and automotive research company that reports market value prices for new and used automobiles. These values are often used by auto dealerships to determine the selling price of their vehicles. The website indicates that the value of a given make and model is normally distributed for cars that are not classified as “antiques”. The average Kelley Blue Book value of a used 2008-2012 sedan is $5,200 with a standard deviation of $985. I would like to determine whether I can ask for a higher selling price in my region for my sedan, so I conduct a random sample of local ads for used 2008-2012 sedans and record the asking price for 28 of them, obtaining an average asking price of $5,485.

(a) Describe the parameter of interest in this study and assign the appropriate symbol.

(b) Write the null and alternative hypotheses for this scenario in both words and symbols.

(c) Determine the values and inequality symbol that should be entered in the One Variable with Sampling applet to create the distribution of simulated sample statistics shown below.

(d) What conclusion should I draw using a significance level of 1%?

(e) What conclusion should I draw using a significance level of 5%?

(f) What conclusion should I draw using a significance level of 10%?

(g) Complete the following statement:

Assuming the average asking price of all 2008-2012 sedans is $5,200, the probability of sampling 28 local adds and finding an average asking price of ____________ or more is ________%.

(h) What asking price should I set for my sedan? Explain your reasoning.
The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all women ages 20-29 to be 5.375 feet (5 feet 4.5 inches) with a standard deviation of 0.283 feet (3.4 inches). You want to determine whether the average height of Dutchess County women ages 20-29 is different than the national average. You survey a random sample of 110 Dutchess County residents ages 20-29. Of the 110 people sampled, 51 identified as women and measured an average height of 5.49 feet (5 feet 5.8 inches).

(a) Describe the parameter of interest in this study and assign the appropriate symbol.

(b) Write the null and alternative hypotheses for this scenario in both words and symbols.

(c) Determine the values and inequality symbol that should be entered in the One Variable with Sampling applet to create the distribution of simulated sample statistics shown below.

(d) What conclusion should we draw using a significance level of 1%?

(e) What conclusion should we draw using a significance level of 5%?

(f) What conclusion should we draw using a significance level of 10%?

(g) Complete the following statement:

Assuming the average height of all women ages 20-29 in Dutchess County is 5.375 feet, the probability sampling 51 Dutchess County women ages 20-29 and measuring an average height of ________ or something more extreme is ________%.
For questions 8 through 14 use the One Proportion or the One Variable with Sampling applet to conduct a statistical analysis for the given scenario by completing parts (a) through (d) below.

(a) Describe the parameter of interest in this study and assign the appropriate symbol.
(b) Write the null and alternative hypotheses for the scenario in both words and symbols.
(c) Determine the values and inequality symbol that should be entered in the applet to create a distribution of simulated sample statistics. Generate the distribution of simulated sample statistics and use the applet to calculate the p-value for the hypotheses in part (b). Include a screenshot of the applet, distribution, and p-value found.
(d) Use the p-value and a significance level of 5% to reach a conclusion. Write your conclusion in the context of the scenario using complete sentences, and appropriate units.

8) A study is regularly conducted by the National Endowment for the Arts to determine music genre preferences in the Hudson Valley over time. The report describes relationships between key demographic characteristics and music preferences. It was noted that in 2010, 24% of adults preferred the genre of Rock/Heavy Metal. In 2020 a random sample of 420 Hudson Valley adults is surveyed with 151 indicating that they prefer Rock/Heavy Metal. Determine whether there is evidence to suggest music preferences have changed in the Hudson Valley in the last 10 years.

9) You were sold a die that is supposed to land on the numbers one and two more frequently than they should occur with a fair die. To determine whether the die is weighted to favor values of one and two you roll the die a total of 67 times and record a result of one or two on 31 of the rolls (36 rolls resulted in a three, four, five, or six).

10) The United States Health and Human Services Poverty Guidelines for 2017 are indicated below:

<table>
<thead>
<tr>
<th>Family/Household Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty Threshold</td>
<td>$12,060</td>
<td>$16,240</td>
<td>$20,420</td>
<td>$24,600</td>
<td>$28,780</td>
<td>$32,960</td>
</tr>
</tbody>
</table>

In March 2017 a PBS report stated that approximately 8.3% of citizens live below the poverty line in the United States. A study is conducted to determine whether a larger proportion of New York State residents live below the poverty line. In a random sample of 3,008 New York State residents, it is determined that 246 were living below the poverty line in 2017.

11) The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years with a standard deviation of 8.5 years, and that life expectancy data is right skewed. After learning this you decide to collect data from a random sample of New York City obituaries from 2018-2020 to compare the life expectancy of New York City residents to all United States citizens to determine whether New York City residents tend to live longer or shorter lives compared to the national average. A random sample of 43 New York City obituaries found an average age at the time of death of 81.1 years.
12) A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest commute time was for workers who live in New York City, who traveled an average of 35.9 minutes each day (or a total of 89.4 hours travel time per year) with a standard deviation of 18 minutes. The data used is right skewed. You live relatively close to the city and know that many of your neighbors work there, so you decide to investigate whether they have a longer average commute time than the average worker who lives in New York City. You survey a total of 36 commuters in your neighborhood and find an average commute time of 41.5 minutes, with a sample standard deviation of 6.1 minutes.

13) A 2017 survey conducted by the National Science Foundation claimed that fewer than 15% of those currently employed in a STEM (Science, Technology, Engineering, and Mathematics) field in the United States are under the age of 30. You conduct a survey to determine whether a lower proportion of STEM field workers in New York State are under the age of 30. After collecting a random sample of 500 STEM field workers in New York State, you find that 67 are under the age of 30.

14) According to their maker, the different colors of Skittles candies (red, orange, yellow, green, and purple) are produced in equal quantities. However, they are packaged by weight and size, not color distribution. Your friend does not believe they are produced equally; rather, a larger proportion of yellow Skittles are produced and packaged. To test this claim you order 10 of the “Fun Size” bags of Skittles from different distribution centers. You then empty all the bags into one bowl and count the number of yellow Skittles. You find that 16 of the 60 Skittles are yellow. Conduct a statistical analysis for the two scenarios below.

Scenario 1: Assuming that Skittles are produced in equal quantities, does this data support your friend’s claim?

Scenario 2: Your friend says they have been recording the proportion of yellow Skittles found in bags for years and they are confident that 1 out of every 3 Skittles produced and packaged is yellow. In this case, let the alternative hypothesis be “the proportion of yellow Skittles produced and packaged is less than one third”.

Follow-Up Question: Compare and contrast your two conclusions. What conclusions are you comfortable drawing about the proportion of yellow Skittles produced and packaged? Explain.
## Section 2.3 | Standardized Statistics and Conclusions in a Hypothesis Test

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fluently use the following vocabulary and symbols:</td>
</tr>
<tr>
<td>o Standardized normal curve</td>
</tr>
<tr>
<td>o Standardized statistic</td>
</tr>
<tr>
<td>o z-score</td>
</tr>
<tr>
<td>o t-score</td>
</tr>
<tr>
<td>• Relate the distribution of simulated sample statistics to the standardized normal curve:</td>
</tr>
<tr>
<td>o Label the distribution with the approximate location of ±1, ±2, and 0 standard deviations from center.</td>
</tr>
<tr>
<td>o Approximate the number of standard deviations from center of a given sample statistic.</td>
</tr>
<tr>
<td>o Approximate the value of the observed statistic that corresponds to a given a number of standard deviations away from the center.</td>
</tr>
<tr>
<td>• Relate a standardized statistic value to the location of an observed statistic on a distribution of sample statistics:</td>
</tr>
<tr>
<td>o Relate the sign of a standardized statistic to the location of the statistic on the distribution.</td>
</tr>
<tr>
<td>o Relate the size of the standardized statistic to the number of standard deviations away from center.</td>
</tr>
<tr>
<td>• Evaluate strength of evidence against the null hypothesis based on a standardized statistic.</td>
</tr>
<tr>
<td>• Understand the relationship between a p-value and the magnitude of a standardized statistic:</td>
</tr>
<tr>
<td>o small p-value ↔ large magnitude (absolute value) standardized statistic</td>
</tr>
<tr>
<td>o big p-value ↔ small magnitude (absolute value) standardized statistic</td>
</tr>
<tr>
<td>• Calculate a standardized statistics using the formula $(\frac{(observed \ statistic) - (mean \ of \ null \ distribution)}{standard \ deviation \ of \ null \ distribution})$</td>
</tr>
</tbody>
</table>

### Key Terms

**Standardized Normal Curve**: The normal distribution for which the mean is 0 and the standard deviation is 1.

**Standardized Statistic (also called Standard Score)**: The standardized statistic for a given data value is the number of standard deviations that the data value is away from the mean of the distribution. Data values above the mean have a positive standard score, data values below the mean have a negative standard score, and the mean of the distribution has a standard score of zero.

**z-score**: The standardized statistic is called the z-score when the variable is categorical and we are considering the simulated distribution of $\hat{p}$ values.

**t-score**: The standardized statistic is called the t-score when the variable is quantitative and we are considering the simulated distribution of $\bar{x}$ values.
Introduction to the Standardized Normal Curve

Recall from Section 1.2 that the standard error is the standard deviation of the distribution of sample statistics. Following are some basic features of the distribution of sample statistics.

- The distribution of sample \( \hat{p} \) values is centered around the true value of \( \pi \) (or the assumed value of \( \pi \) in the case of a hypothesis test).
- The distribution of sample \( \bar{x} \) values is centered around the true value of \( \mu \) (or the assumed value of \( \mu \) in the case of a hypothesis test).
- This distribution of sample \( \hat{p} \) values will be roughly normally distributed for even a relatively small sample size.
- The distribution of sample \( \bar{x} \) values is called the “t-distribution”. The t-distribution is roughly bell-shaped and becomes closer and closer to a normal distribution as the sample size increases. The t-distribution of \( \bar{x} \) values will be roughly normally distributed when the sample size is larger than 30 or when the population itself has data that is roughly normally distributed.

Recall that we can estimate the standard deviation of a normal distribution by comparing the simulated distribution of sample statistics to the standardized normal distribution (pictured below) in order to approximate the locations in the distribution where we would find sample statistics that are 1 standard deviation, 2 standard deviations, or 3 standard deviations above or below the mean. The picture below indicates the location of the sample statistics that are in the middle (at the mean), 1 standard deviation above/below the mean, and 2 standard deviations above/below the mean.
The **standardized normal curve** is the normal distribution for which the mean is 0 and the standard deviation is 1. Positive values are above the mean, and negative values are below the mean. The value along the horizontal indicates the number of standard deviations above or below the mean. Therefore, the standardized normal curve looks as follows:

![Standardized Normal Curve](image)

We can take any normal distribution of sample statistics and correlate it to the standardized normal curve by noting that the **standardized normal curve indicates how many standard deviations each sample statistic is away from the mean of the distribution.**
There are three candy bars on the market that are in direct competition: Yummies, Crunchies, and Tasties. Yummies wants to know whether they got more than 40% of all these candy bar purchases during the previous year. The hypotheses are as follows: \( H_0: \pi = 0.4 \) and \( H_A: \pi > 0.4 \), where \( \pi \) is the proportion of all of these candy bar purchases that were Yummies purchases. **The distribution of simulated sample statistics below has a mean of 0.40 and a standard deviation of 0.05.** Label the distribution to clearly correlate the distribution with the standardized normal distribution.

**Solution 2.3.1:**

In this distribution of simulated sample statistics, which could also be called a distribution of simulated \( \hat{p} \) values, the mean is 0.4, which correlates with “0 standard deviations above/below the mean” in the standardized normal distribution.

- 0 standard deviations above/below the mean is the simulated statistic 0.4 (labeled as 0).
- The mean is 0.4 and the standard deviation is 0.05.
  - 1 standard deviation above the mean is the simulated statistic 0.4+0.05 = 0.45 (labeled as 1).
  - 1 standard deviation below the mean is the simulated statistic 0.4-0.05 = 0.35 (labeled as -1).
- The mean is 0.4 and the standard deviation is 0.05.
  - 2 standard deviations above the mean is the simulated statistic 0.4+0.05+0.05 = 0.50 (labeled as 2).
  - 2 standard deviation below the mean is the simulated statistic 0.4-0.05-0.05 = 0.30 (labeled as -2).
Suppose the average (mean) GPA for all students at a certain college is 2.74. Suppose we collect a random sample of 85 students at the school and record the average (mean) GPA of the group of 85 students. The distribution of simulated sample statistics below has a mean of 2.74 and a standard deviation 0.02. The distribution has been labeled to correlate it with the standardized normal curve. Identify the values of the simulated statistics that correspond to each of the labeled values (-2, -1, 0, 1, 2).

Solution 2.3.2

In the simulated distribution, which could also be called a distribution of \( \bar{x} \) values, the mean is 2.74, which corresponds to “0 standard deviations above/below the mean” in the standardized normal distribution. Thus, the middle value that is labeled as 0 is re-labeled as 2.74.

- 0 standard deviations above/below the mean is the simulated statistic 2.74. The value that is labeled as 0 corresponds to a mean GPA of 2.74 from a random sample of 85 students.
- 1 standard deviation above the mean is the simulated statistic 2.74 + 0.02 = 2.76. The value that is labeled as 1 corresponds to a mean GPA of 2.76 from a random sample of 85 students.
- 2 standard deviations above the mean is the simulated statistic 2.74 + 0.02 + .02 = 2.78. The value that is labeled as 2 corresponds to a mean GPA of 2.78 from a random sample of 85 students.
- 1 standard deviation below the mean is the simulated statistic 2.74 − 0.02 = 2.72. The value that is labeled as −1 corresponds to a mean GPA of 2.72 from a random sample of 85 students.
- 2 standard deviations below the mean is the simulated statistic 2.74 − 0.02 − .02 = 2.70. The value that is labeled as −2 corresponds to a mean GPA of 2.70 from a random sample of 85 students.
According to a recent article in the Washington Post¹, the average (mean) screen time for 8-to-12-year-olds in the U.S. is 284 minutes per day (which is 4 hours and 44 minutes per day). Suppose we are interested in determining whether the 8-to-12-year-olds in our area have the same average screen time, so we collect a random sample of 104 kids in our area who are 8 to 12 years old.

The distribution of simulated sample statistics below has a mean of 284 and a standard deviation of 3.

Work together to complete the following. Be ready to share with the class.
Label the given distribution to correlate the distribution with the standardized normal distribution, including adding vertical lines and labeling those lines with the values -2, -1, 0, 1, and 2, and also with the values of the sample means that correspond to the vertical lines. For example, there should be a vertical line labeled “0 standard deviations above/below the mean”, which corresponds to a sample mean value of $\bar{x} = 284$.

Samaya suspects that the die in her board game is rolling the number five more often than it should if the die was fair. She makes the following hypotheses: $H_0: \pi = \frac{1}{6} \approx 0.167$ and $H_A: \pi > \frac{1}{6} \approx 0.167$, where $\pi$ is the proportion of times her die rolls the number five in the long run. Samaya decides that she will roll the die 60 times to see if it rolls a five significantly more often than it would if it were a fair die.

The distribution of simulated sample statistics below has a mean of 0.167 and a standard deviation of 0.05.

Work together to answer the following questions. Be ready to share with the class.
Label the given distribution to correlate the distribution with the standardized normal distribution, including adding vertical lines and labeling those lines with the values -2, -1, 0, 1, and 2, and also with the sample proportion values that correspond to the vertical lines. For example, there should be a vertical line labeled “0 standard deviations above/below the mean”, which corresponds to a sample proportion value of $\hat{p} = 0.167$.

![Proportion of successes](https://example.com/distribution.png)

- proportion of times (out of 60 rolls) that a five is rolled when the die is fair
Introduction to the Standardized Statistic

- The **standardized statistic**, also called the **standard score**, tells us how many standard deviations an observed statistic is from the mean of the null distribution.
  - A positive standardized statistic (standard score) means the statistic is higher than the mean of the null distribution.
  - A negative standardized statistic (standard score) means the statistic is lower than the mean of the null distribution.
  - Example (see distribution below): A standardized statistic of 1.4 means that the observed statistic is 1.4 standard deviations above the mean of the null distribution.
  - Example (see distribution below): A standard score of $-0.8$ means that the observed statistic is 0.8 standard deviations below the mean of the null distribution.

- As the magnitude (absolute value) of the standardized statistic gets bigger, the evidence against the null hypothesis gets stronger and the p-value gets smaller.

- The bigger the magnitude (absolute value) of the standard score, the further the observed statistic is from the center of the null distribution.
  - A standard score bigger than 2 or smaller than $-2$ means that the observed statistic is unusual/unexpected according to the 2 Standard Deviations Rule, and this would be associated with a "small" p-value.
  - A standard score between $-2$ and 2 means that the observed statistic is usual/expected according to the 2 Standard Deviations Rule, and this would be associated with a "big" p-value.
There are three candy bars on the market that are in direct competition: Yummies, Crunchies, and Tasties. Yummies wants to know whether they got more than 40% of all these candy bar purchases during the previous year. The distribution of simulated sample statistics below, which could also be called the distribution of $\hat{p}$ values, has a mean of 0.40 and a standard deviation of 0.05.

![Distribution of simulated sample statistics](image)

**Example 2.3.3**

There are three candy bars on the market that are in direct competition: Yummies, Crunchies, and Tasties. Yummies wants to know whether they got more than 40% of all these candy bar purchases during the previous year. The distribution of simulated sample statistics below, which could also be called the distribution of $\hat{p}$ values, has a mean of 0.40 and a standard deviation of 0.05.

**Solution 2.3.3:**

(a) If the observed sample statistic was $\hat{p} = 0.425$, then what is the value of the standardized statistic. Would this be associated with a big $p$-value or a small $p$-value? What conclusion would we make about the hypotheses?

(b) If the standard score for the observed statistic was 2.5, then what is the value of the observed statistic from the sample? Would this be associated with a big $p$-value or a small $p$-value? What conclusion would we make about the hypotheses?
Calculating the Standardized Statistic using a Formula

The standardized statistic (standard score) value tells us the number of standard deviations that the observed statistic is above or below the center (mean) of the null distribution. Positive standard scores correspond to statistics that are above the mean, while negative standard scores correspond to statistics that are below the mean. The formula that can be used to calculate the value of the standardized statistic (standard score) is shown below.

**Formula for Standardized Statistic**

\[
\text{standardized statistic} = \frac{(\text{observed statistic}) - (\text{mean of null distribution})}{(\text{standard deviation of null distribution})}
\]

Notice that this formula directs us to calculate the distance that the observed statistic is from the mean of the null distribution. The numerator will come out negative if the observed statistic is below the mean and the numerator will come out positive if the observed statistic is above the mean. The numerator is then divided by the standard deviation of the null distribution (also called the standard error) in order to tell us “how many standard deviations” the statistic is from the center (mean) of the null distribution.

**COMMON ERRORS ALERT**

- **It is very common for students to accidentally reverse the order of the subtracted values in the numerator.**
  Be careful to not make this error! Making this error will result in your answer having the wrong sign! To avoid this error, be sure to visualize the null distribution and consider where the observed statistic is within that distribution. If the observed statistic is above the mean of the null distribution, then your standard score calculation should have an answer that is positive. If the observed statistic is below the mean of the null distribution, then your standard score calculation should have an answer that is negative.
  **Example:** Suppose the distribution of sample statistics is centered at 150 with a standard error of 10, and suppose the observed statistic is 135.
  Incorrect method: \[\frac{150 - 135}{10} = \frac{15}{10} = 1.5\] We know this is wrong since the observed statistic is BELOW the center (mean) of the distribution. So, the standard score should be negative.
  Correct method: \[\frac{135 - 150}{10} = \frac{-15}{10} = -1.5\] This is the correct answer. The observed statistic is 1.5 standard deviations below the center (mean) of the distribution.

- **It is very common for students to accidentally type the calculation into the calculator incorrectly.**
  Be careful to not make this error! To avoid this error, be sure to visualize the null distribution and consider approximately how many standard deviations the observed statistic is above or below the center (mean) of the distribution. If you get an answer in the calculator that is a lot different than your approximation, then you have likely made an error in the calculator.
  **Example:** Calculate the value \(\frac{135 - 150}{10}\).
  Incorrect method:

  \[
  135 \cdot 150 \div 10 = 120
  \]
This answer would indicate that the observed statistic is 120 standard deviations above the mean of the null distribution, which is clearly incorrect. Notice that the mean is 150, the standard deviation is 10, and the observed statistic is 135, which is just a little below the mean of the null distribution (not 120 standard deviations above the mean). The error here is that we entered the expression in the calculator in a manner that will divide the 150 by 10 instead of dividing the entire numerator (135-150) by 10.

Correct method:

\[
\frac{(135 - 150)}{10} = -1.5
\]

This is the correct answer. The observed statistic is 1.5 standard deviations below the mean of the null distribution. This time, we used parentheses around the numerator in order to indicate to the calculator that the entire numerator should be divided by 10.
Example 2.3.4

Suppose that the proportion of success in the population is 40% and suppose we create a distribution of sample statistics (distribution of \(\hat{p}\) values) that has a standard deviation of 0.03. Also suppose that the observed statistic is 0.352. Based on this information, complete the following.

(a) Sketch and label a distribution with the values of the sample statistics along the horizontal.
(b) Label your distribution with the standardized statistic values along the horizontal.
(c) Identify whether the standard score of the observed statistic is positive or negative or 0 based on the labeled distribution created in parts (a) and (b). Estimate the value of the standard score.
(d) Identify the correct symbol (t or z) for the standard score.
(e) Calculate the value of the standardized statistic using the formula.

Solution 2.3.4:

(a) and (b) The proportion of successes in the population is 40% = 0.4, which is the center of the distribution of sample statistics. The distribution has a standard deviation (called the standard error) of 0.03. The sample statistic and standard score values are labeled on the distribution below.

(c) The observed statistic is \(\hat{p} = 0.352\). This value would fall between 0.34 and 0.37, which means it would fall between a standard score of \(-2\) and \(-1\). An estimate of the standard score may be a value near \(-1.5\).

(d) The standard score will be called a z-score since this is a distribution of \(\hat{p}\) values.

(e) The z-score is \[ z = \frac{0.352 - 0.4}{0.03} = \frac{-0.048}{0.03} = -1.6 \]. This means that the observed statistic, \(\hat{p} = 0.352\), is 1.6 standard deviations below the center (mean) of the distribution. This makes sense based on the labeled sketch of the distribution of sample statistics. The observed statistic would be located approximately HERE.
Example 2.3.5

Suppose that the mean of a population is 450, that the distribution of sample statistics has a standard deviation of 8, and that the observed statistic of \( \bar{x} = 456.2 \). Based on this information, do the following:

(a) Sketch and label a distribution with the values of the sample statistics along the horizontal.

(b) Label the distribution with the standardized statistic values along the horizontal.

(c) Identify whether the standard score of the observed statistic is positive, negative, or 0 based on the labeled distribution created in parts (a) and (b). Estimate the value of the standard score.

(d) Identify the correct symbol (t or z) for the standard score.

(e) Calculate the value of the standardized statistic using the formula.

Solution 2.3.5:

(a) and (b) The population mean is 450, which is the center of the distribution of sample statistics. The distribution has a standard deviation (called the standard error) of 8. The sample statistic and standard score values are labeled on the distribution below.

(c) The observed statistic is \( \bar{x} = 456.2 \). This value would fall between 450 and 458, which means it would fall between a standard score of 0 and 1. An estimate of the standard score may be a value near 0.5.

(d) The standard score will be called a t-score since this is a distribution of \( \bar{x} \) values.

(e) The t-score is \( t = \frac{456.2 - 450}{8} = \frac{6.2}{8} = 0.8 \). This means that the observed statistic, \( \bar{x} = 456.2 \), is 0.8 standard deviations above the center (mean) of the null distribution. This makes sense based on the labeled sketch of the distribution of sample statistics. The observed statistic would be located approximately HERE.

Careful!

It is a common error for students who are learning statistical concepts to confuse the following terms:

- Standard deviation of a data set
- Standard error (standard deviation of the distribution of sample statistics)
- Standardized statistic
- Observed statistic

Are you able to define the terms above, give the appropriate symbols that are used for those terms, and explain the relationships that exist between the terms?
Suppose Politician A has a 48% approval rating in the general electorate. We select a random sample of 250 people in the general electorate and ask them whether they approve of Politician A.

Suppose that the observed statistic is \( \hat{p} = 110/250 = 44\% \) who approve of Politician A. We create a distribution of sample statistics based on the knowledge that 48% of the total electorate approves of Politician A, and this distribution has a standard deviation of 0.03.

Based on the above scenario, answer the following questions in groups and be ready to share with the class.

(a) Label the distribution below with the values of \( \hat{p} \) matching each vertical line.
(b) Label the distribution below with the standardized statistic values matching each vertical line.
(c) Identify whether the standard score of the observed statistic is positive, negative, or 0 based on the labeled distribution from parts (a) and (b). Estimate the value of the standard score.
(d) Identify the correct symbol (t or z) for the standard score.
(e) Calculate the value of the standardized statistic using the formula.

(sample statistic values, p-hats)

(standard score values)
In recent years, the average (mean) time for students to earn a 4-year degree is 76 months (which is 6 years and 4 months). Suppose we collect a random sample of 85 college graduates from recent years and calculate the average (mean) amount of time that the sample of graduates took to earn a 4-year degree. The observed statistic is $\bar{x} = 52.1$. We create a distribution of simulated sample statistics based on the assumption that there is a mean degree time of 76 months. The distribution of simulated sample statistics has a standard deviation of 0.08.

Based on the above scenario, answer the following questions in groups and be ready to share with the class.

(a) Label the distribution below with the values of $\bar{x}$ matching each vertical line.
(b) Label the distribution below with the standardized statistic values matching each vertical line.
(c) Identify whether the standard score of the observed statistic is positive, negative, or 0 based on the labeled distribution from parts (a) and (b). Estimate the value of the standard score.
(d) Identify the correct symbol (t or z) for the standard score.
(e) Calculate the value of the standardized statistic using the formula.

1. [Link to PDF](http://nces.ed.gov/pubs2013/2013150.pdf)
Section 2.3 | Standardized Statistics and Conclusions in a Hypothesis Test

Exercises

1) Suppose the hypotheses in a statistical study are as follows:
   \[ H_0: \pi = 0.25 \]
   \[ H_a: \pi < 0.25 \]
   In the context of these hypotheses, which of the following standardized statistics would provide the strongest evidence against the null hypothesis? Explain your reasoning.
   
   \[ z = -1 \quad z = 0 \quad z = 3 \quad z = -1.80 \]

2) (Multiple Choice) Suppose the hypotheses in a statistical study are as follows:
   \[ H_0: \pi = 0.25 \]
   \[ H_a: \pi < 0.25 \]
   In the context of these hypotheses, which of the following standardized statistics would provide the strongest evidence against the null hypothesis?
   (a) \( z = -1 \)
   (b) \( z = 0 \)
   (c) \( z = 3 \)
   (d) \( z = -1.80 \)

3) (True or False) A small (close to 0) p-value goes with a large magnitude (far from 0) standardized statistic.
   (a) True
   (b) False

4) Consider the screenshot below that was obtained using the One proportion inference applet. Use information from the screenshot to find the standardized statistic for a sample proportion value of \( \hat{p} = 0.44 \).
5) Josh is a foosball player who is working on his trick shots. While practicing, he gets 12 out of 20, or 60%, of his trick shots in the goal. Josh wants to know if his long-run proportion of making his trick shots is greater than 50%.

(a) What is the value of the observed statistic? Use the appropriate symbol in your answer.

(b) Suppose you use the **One proportion inference** applet to generate 100 simulated sample statistics. The resulting dot plot is shown below. Fill in the blanks on the left that were used in the applet.

![Dot plot](image)

(c) What is the value of the p-value in this case? Circle the dots in the dot plot that correspond to the p-value. How would you interpret the p-value in terms of Josh’s ability to make trick shots?

(d) Calculate the standardized statistic, including the appropriate symbol. Does the strength of evidence associated with the standard score match the strength of evidence associated with the p-value from part (c)? Explain.

6) A 2017 survey conducted by the National Science Foundation claimed that fewer than 15% of those currently employed in a STEM field in the United States of America are under the age of 30. You conduct a similar survey to determine whether the proportion of STEM field workers in your state under the age of 30 is less than the national average. After collecting a random sample of 500 STEM field workers in your state, you find that 12% of those surveyed are under the age of 30.

(a) Will the standardized statistic for your sample statistic be positive or negative? Explain.

(b) Suppose you generate a distribution of sample statistics that has a standard deviation of 0.02. Sketch the distribution using a normal curve and label the distribution with the parameter, statistic, and the locations of the standard score values of -2, -1, 0, 1, and 2.

![Normal curve](image)

(c) Calculate the standardized statistic and compare your result to your statement in part (a) and the position of your statistic in part (b).

(d) What conclusion would you draw from your z-score and an alternative hypothesis of \( H_A: \pi < 0.15 \)?
7) Every Tuesday morning Sara, Gary, and Josh gather to determine who is the “luckiest” that week, then they send that person to purchase a lottery ticket. They each take 5 minutes to flip a coin as many times as possible, recording the total number of heads. The results for last week are recorded in the table below:

<table>
<thead>
<tr>
<th></th>
<th>Sara</th>
<th>Gary</th>
<th>Josh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heads</td>
<td>115</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>270</td>
<td>54</td>
<td>108</td>
</tr>
</tbody>
</table>

a) Calculate the sample statistics for Sara, Gary, and Josh. Who would you claim was the “luckiest” according to the sample statistics? Explain.

b) Calculate the standardized statistic for Sara, Gary, and Josh assuming the distributions of sample statistics have standard deviations of 0.03, 0.069, and 0.048 respectively. Who would you claim was the “luckiest” according to the standardized statistics? Explain.

8) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. For one week PepsiCo surveyed 381 people and had 201 people choose Pepsi. Can PepsiCo claim that the majority of people prefer Pepsi over Coke?

a) What is the parameter of interest in the context of this study (in words)? What symbol should be used to represent the parameter?

b) State the null and alternative hypotheses using the symbol defined above.

c) What is the observed statistic and what symbol should be used to represent it?

a) Use the One proportion inference applet to generate a distribution of simulated sample statistics and use the distribution to calculate the standardized statistic. Then use the values provided by your distribution to label the mean ($\pi$), sample statistic, and sample statistic values for the standard scores displayed below.

![Distribution of simulated sample statistics]

---

$(sample\ statistic\ values, p-hats)$

$-2 \ -1 \ 0 \ 1 \ 2 \ (standard\ score\ values)$

---

d) What conclusion would you draw from your z-score regarding PepsiCo’s claim?
9) A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest average commute was for workers in New York City, who traveled an average of 35.9 minutes each day. You live relatively close to New York City and know of several residents in your neighborhood who commute into the city each day. You decide to investigate whether the commuters in your neighborhood have a different average commute time than workers in New York City. You survey 25 commuters in your neighborhood and find an average commute time of 41.2 minutes, with a standard deviation of 12.9 minutes.

a) Write out the null and alternative hypotheses for this scenario in words and symbols.

b) Label the mean (\( \mu \)), sample statistic, and sample statistic values for the standard scores displayed.

\[
\begin{array}{cccc}
-2 & -1 & 0 & 1 \\
\end{array}
\]

\( \text{(sample statistic values, x-bars)} \)

\[
\begin{array}{cccc}
-2 & -1 & 0 & 1 \\
\end{array}
\]

\( \text{(standard score values)} \)

c) Estimate your standard score (t-score) using the distribution from part (b).

d) Calculate the t-score and compare it to your estimate in part (c).

e) What conclusion would you draw from your t-score regarding your research question?
10) While investigating online, you read a statement claiming that the average height of women between ages 20 and 40 is 65 inches. To test the validity of this claim, you decide to collect a random sample of 74 women in your town between the ages of 20 and 40. You find that the average height of women in your sample to be 67.5 inches, with a sample standard deviation of 3.74 inches.

a) Write out the null and alternative hypotheses in symbols and words for the research question, “Is the average height of women in your town ages 20 to 40 greater than 65 inches?”

b) Label the mean, sample statistic, and sample statistic values for the standard scores displayed below.

![Histogram of sample means]

\[ \begin{array}{cccccc}
 \text{Sample statistic values, } x-bar & -2 & -1 & 0 & 1 & 2 \\
 \text{Standard score values} & & & & & \\
\end{array} \]

(c) Estimate your t-score based on the position of the sample statistic in the above distribution, then calculate the t-score.

d) Based on your t-score, write out a conclusion in context regarding your research question.

11) A study was conducted by the National Endowment for the Arts (C02-91) to described music preferences in the U.S. and how they have changed between 1982 and 2002. The report describes relationships between key demographic characteristics and music preferences. In this report, it was noted that in 2002, the genre of Rock/Heavy Metal (e.g. Nirvana, The Clash, R.E.M) had a fan base of over 50 million adults nationwide (24% of the population), with the first group of millennials (a.k.a. “The Net Generation”) making up 46% of this group. A new survey is conducted in 2020 in the Hudson Valley to determine whether these proportions have changed. A random sample of 180 Hudson Valley adults found that 56 prefer Rock/Heavy Metal.

a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.

b) Sketch the graph of the distribution and identify the values of the simulated statistics that are -2, -1, 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.

c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.

d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.

e) Write a complete sentence to give a conclusion in context.
12) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that people are more likely to prefer their brand of cola to Coca-Cola. A random sample of 315 people is taken, and 53% of them choose Pepsi as their preferred cola.

a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.

b) Sketch the graph of the distribution and identify the values of the simulated statistics that are -2, -1, 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.

c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.

d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.

e) Write a complete sentence to give a conclusion in context.

13) The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years with a standard deviation of 6.5 years. After learning this you decide to collect data from a random sample of obituaries of Dutchess County residents from 2018-2020 to compare the life expectancy of Dutchess County residents to all United States citizens. A random sample of 39 local obituaries found an average age at the time of death of 79.1 years.

a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.

b) Sketch the graph of the distribution and identify the values of the simulated statistics that are -2, -1, 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.

c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.

d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.

e) Write a complete sentence to give a conclusion in context.

14) A migraine is a particularly painful type of headache that patients sometimes wish to treat with acupuncture. An advertisement for an acupuncturist in the Hudson Valley claims that 1 in 3 of their clients reports relief from regular migraines for at least 24 hours after an acupuncture treatment at their facility. To determine whether acupuncture relieves migraine pain, researchers conducted a study where 89 New York State residents who were experiencing migraine headaches were either given acupuncture treatment designed to treat migraines, or they were given a placebo treatment (using acupuncture, but in areas not intended to treat migraines). After 24 hours patients were asked if they were now “pain-free”. Of the 89 participants, 43 received acupuncture intended to relieve migraines, and 46 did not. Of the 43 people who received the acupuncture treatment, 19 indicated they were pain-free after 24 hours. Of the 46 people who received the “false” treatment, 2 indicated they were pain-free after 24 hours.

a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.

b) Sketch the graph of the distribution and identify the values of the simulated statistics that are -2, -1, 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.
c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.
d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.
e) Write a complete sentence to give a conclusion in context.

15) You conduct a survey to determine whether fewer than 15% of STEM field workers in your state are under the age of 30. After collecting a random sample of 210 STEM field workers in your state, you find that 25 are under the age of 30.
a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.
b) Sketch the graph of the distribution and identify the values of the simulated statistics that are \(-2\), \(-1\), 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.
c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.
d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.
e) Write a complete sentence to give a conclusion in context.

16) A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest average commute was for workers in New York City, who traveled an average of 35.9 minutes each day with a standard deviation of 9.8 minutes. You live relatively close to New York City and know of several residents in your neighborhood who commute into the city each day. You decide to investigate whether the commuters in your neighborhood have a longer average commute time than workers in New York City. You survey 21 commuters in your neighborhood and find an average commute time of 39.2 minutes, with a standard deviation of 10.3 minutes.
a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.
b) Sketch the graph of the distribution and identify the values of the simulated statistics that are \(-2\), \(-1\), 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.
c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.
d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.
e) Write a complete sentence to give a conclusion in context.
17) The United States Health and Human Services Poverty Guidelines for the 48 contiguous states and the District of Columbia in 2019 are displayed below:

According to census.gov, approximately 10.5% of citizens were living below the poverty line in the United States in 2019. A study is conducted to determine whether that percentage is representative of New Hampshire residents in 2020. In a random sample of 2,580 New Hampshire residents, it was found that 178 were living below the poverty line.

a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.

b) Sketch the graph of the distribution and identify the values of the simulated statistics that are -2, -1, 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.

c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.

d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.

e) Write a complete sentence to give a conclusion in context.

18) The United States Health and Human Services Poverty Guidelines for the 48 contiguous states and the District of Columbia in 2019 are displayed below:

According to census.gov, approximately 10.5% of citizens were living below the poverty line in the United States in 2019. A study is conducted to determine whether that percentage is representative of Mississippi residents in 2020. In a random sample of 1,843 Mississippi residents it was found that 361 were living below the poverty line.

a) Use the One proportion inference or the One Variable with Sampling applet to generate the necessary values to calculate the standardized statistic. State the values entered in the applet to generate the distribution of simulated statistics.

b) Sketch the graph of the distribution and identify the values of the simulated statistics that are -2, -1, 0, 1, and 2 standard deviations away from the mean. Estimate the standardized statistic using this graph.

c) Calculate the standardized statistic using the values from part (a) and include the appropriate symbol.

d) Use the standardized statistic to make a statement about whether the observed statistic is usual or unusual.

e) Write a complete sentence to give a conclusion in context.

19) Below are two follow-up questions regarding questions 17 and 18. It appears that New Hampshire and Mississippi are on opposite ends of the spectrum regarding poverty rates.

(a) Are the observed statistics from questions 17 and 18 fair to compare side-by-side? Why or why not?

(b) Can you think of any other variables that might be relevant to this discussion? Explain.
Strength of Evidence Review

Recall that a **smaller p-value gives stronger evidence against the null hypothesis**. The p-value tells us the probability of getting a sample like the one we observed or more extreme if the null hypothesis is true. Since the distribution of sample statistics shows us what the sample statistics are expected to look like if the null hypothesis is true, we can conclude that a smaller p-value indicates that the observed sample statistic is further in the tail of the distribution. This makes the observed sample statistic less probable if the null hypothesis is true, and so there is stronger evidence against the null hypothesis.

Recall that a **larger magnitude (absolute value) standardized statistic gives stronger evidence against the null hypothesis**. The standardized statistic (z-score or t-score) tells us the number of standard deviations that the observed statistic is away from the center of the null distribution (assumed value of the population parameter). Since the distribution of sample statistics shows us what the sample statistics are expected to look like if the null hypothesis is true, we can conclude that a larger magnitude (absolute value) standardized statistic indicates that the observed sample statistic is further in the tail of the distribution. This makes the observed sample statistic less probable if the null hypothesis is true, and so there is stronger evidence against the null hypothesis.
Strength of Evidence and the Location of the Observed Statistic

Holding all other things constant, the further an observed statistic is in the tail of a distribution of sample statistics, the stronger the evidence against the null hypothesis.

- Holding all other things constant, an observed statistic that is further in the tail of the null distribution will create a smaller p-value.
- Holding all other things constant, an observed statistic that is further in the tail of the null distribution will create a larger magnitude (absolute value) standardized statistic.

Example 2.4.1

Politicians want to understand the voting population’s support for a new tax. A previous poll concluded that 35% of voters supported the new tax, and the polling company wants to know whether that 35% support is still true.

- Basheer takes a random sample of 150 registered voters and finds that 58 of them support the new tax.
- Fatima takes a random sample of 150 registered voters and finds that 40 of them support the new tax.

Who will have stronger evidence against the null hypothesis based on his/her sample? Explain.

Solution 2.4.1:

In order to answer this question, we observe the following:

- Both Basheer and Fatima have a null hypothesis that $H_0: \pi = 0.35$ and an alternative hypothesis that $H_A: \pi \neq 0.35$.
- Both Basheer and Fatima have a sample size of n = 150.
- Because they have the same null hypothesis and the same sample size, Basheer and Fatima have approximately the same distribution of simulated sample statistics (all simulations come out slightly different).
- Basheer’s sample statistic is $\hat{p} = \frac{58}{150} = 0.3867$, which is a distance of 5.2 units above the center (52.5) of the null distribution.
- Fatima’s sample statistic is $\hat{p} = \frac{40}{150} = 0.267$, which is a distance of 12.5 units below the center (52.5) of the null distribution.

All things are held constant except for the fact that Fatima has an observed statistic that is further in the tail of the distribution. Thus, Fatima has stronger evidence against the null hypothesis. Fatima’s sample statistic provides strong evidence that the proportion of voters who support the new tax is not 35%.
### Strength of Evidence and the Sample Size

Holding all other things constant, **the larger the sample size, the stronger the evidence** against the null hypothesis.

- Holding all other things constant, the distribution of sample statistics will become more and more narrow (have a smaller standard error) as the sample size increases, which means the observed statistic will be further in the tail of this narrower distribution.
- Holding all other things constant, a larger sample size will correspond to a smaller p-value.
- Holding all other things constant, a larger sample size will correspond to a larger magnitude (absolute value) standardized statistic.

### Example 2.4.2

Politicians want to understand the voting population’s support for a new tax. A previous poll concluded that 35% of voters supported the new tax, and the polling company wants to know whether that 35% support is still true.

- Basheer takes a random sample of 150 registered voters and finds that 45 of them support the new tax; \( \hat{p}_{\text{Basheer}} = \frac{45}{150} = 0.30 = 30\% \).
- Fatima takes a random sample of 1500 registered voters and finds that 450 of them support the new tax; \( \hat{p}_{\text{Fatima}} = \frac{450}{1500} = 0.30 = 30\% \).

Who will have stronger evidence against the null hypothesis based on his/her sample? Explain.

#### Solution 2.4.2:

In order to answer this question, we observe the following:

- Both Basheer and Fatima have a null hypothesis that \( H_0: \pi = 0.35 \) and an alternative hypothesis that \( H_A: \pi \neq 0.35 \).
- Basheer and Fatima each have the same observed sample statistic value of \( \hat{p} = 0.30 \).
- The only difference between the samples is the sample size.

All things are held constant except for the sample size. Since Fatima has the larger sample size, **Fatima’s sample statistic provides stronger evidence against the null hypothesis**. Fatima’s observed statistic provides stronger evidence that the proportion of voters who support the new tax is not 35%.

To help visualize what is happening, a distribution of simulated sample statistics based on each sample size is shown below:

(continued on next page)
Solution 2.4.2 (continued):
The simulated distributions for both Basheer and Fatima are given below using the same scale along the horizontal axis.

Notice that the distribution of sample statistics is narrower for the larger sample. For Basheer, the distribution of sample statistics shows that it would be usual and expected to get sample statistics between 0.27 and 0.43. For Fatima, the distribution of sample statistics shows that it would be usual and expected to get sample statistics between 0.33 and 0.37.

The observed statistic that they each observe, \( \hat{p} = 0.30 \), is further in the tail of the distribution based on the larger sample size. The observed sample statistic is unusual/unexpected within Fatima’s distribution, but the observed sample statistic is usual/expected within Basheer’s distribution.
Strength of Evidence and One-Sided vs. Two-Sided Testing

Holding all other things constant, a one-sided test results in stronger evidence against the null hypothesis.

- Holding all other things constant, a one-sided test will have a p-value that is half the size of the p-value from a two-sided test (in a simulation, this is approximately half the size).
- Holding all other things constant, a two-sided test will have a p-value that is twice the size of the p-value from a one-sided test (in a simulation, this is approximately twice the size).

Example 2.4.3

Politicians want to understand the voting population’s support for a new tax. A previous poll concluded that 35% of voters supported the new tax.

- Basheer wants to know whether registered voters support the new tax at a rate different than 35%. The alternative hypothesis is $H_A: \pi \neq 0.35$, “The proportion of all registered voters that support the new tax is not 35%”. Basheer takes a random sample of 150 registered voters and finds that 42 of them support the new tax, so $\hat{p} = \frac{42}{150} = 0.28 = 28\%$.

- Fatima wants to know whether fewer than 35% of registered voters support the new tax. The alternative hypothesis is $H_A: \pi < 0.35$, “The proportion of all registered voters that support the new tax is less than 35%”. Fatima takes a random sample of 150 registered voters and finds that 42 of them support the new tax, so $\hat{p} = \frac{42}{150} = 0.28 = 28\%$.

Who will have stronger evidence against the null hypothesis based on his/her alternative hypothesis? Explain.

Solution 2.4.3:

In order to answer this question, we need to observe the following:

All things are held constant in the two scenarios except for the alternative hypothesis. Phrasing the alternative hypothesis as $H_A: \pi \neq 0.35$ is using a two-sided test, and phrasing the alternative hypothesis as $H_A: \pi < 0.35$ is using a one-sided test. The one-sided alternative hypothesis, $H_A: \pi < 0.35$, results in stronger evidence against the hypothesis that 35% of registered voters support the tax. The two-sided test results in a p-value that is about twice the size of the p-value in the one-sided distribution because we are counting dots in both tails instead of only one tail.
Reasoning Behind Using Two-Sided instead of One-Sided Test

The discerning student may, at this time, question why researchers don’t always phrase research questions in a manner that would create one-sided tests, which would therefore result in stronger evidence against the null hypothesis than a two-sided test. After all, stronger evidence is better than weaker evidence, right?

Well, not exactly. There are two main concepts to understand in relation to this question.

1. **It is important to maintain impartiality.**
   It is important for researchers to always maintain as much objectivity/impartiality/neutrality as possible in order to maintain the most integrity and validity in the statistical analysis process. If the researcher makes a conjecture that the population parameter is bigger/smaller than a specific value in the first step of the hypothesis testing procedure, then the researcher would be starting the hypothesis testing process by already excluding a huge set of values from consideration. That is not impartial.
   For example, if we created a null hypothesis that \( \mu = 200 \) and an alternative hypothesis that \( \mu < 200 \), then we have already excluded the possibility that \( \mu > 200 \) in our hypothesis testing process, and we would only be counting sample statistics that are less than or equal to the observed statistic when calculating our p-value. It is only our conjecture made in the alternative hypothesis that leads us to count in this manner. This is a conjecture that may be incorrect! It would be better to create an alternative hypothesis that more broadly captures the possibilities for the true value of the population parameter. It would be more impartial to have the alternative hypothesis be \( \mu \neq 200 \).

2. **It is better to have a higher, more difficult standard to achieve “enough evidence”.**
   Simply put, it is more difficult to find enough evidence to reject the null hypothesis if we use a two-sided test. Recall that, holding all other things constant, a two-sided test will create a p-value that is about twice the size of the p-value from the corresponding one-sided test. It is more difficult to get a p-value that is smaller than the significance level, \( \alpha \), if we are counting twice the number of sample statistics. Researchers are being held to a higher standard when using a two-sided test since, in order for the p-value to be less than \( \alpha \), the observed statistic must be further in the tail of the null distribution to achieve statistical significance.
Example 2.4.4

Suppose a drug company creates a new drug that has life-saving properties, but also has a debilitating side-effect that impacts a small percentage of people. Before releasing the drug for sale to save people’s lives, many studies are conducted on the side-effects of the drug in order to ensure that the there is a true understanding of the percentage of people who will be impacted by this debilitating side-effect. Once researchers truly understand the percentage of people who are expected to experience the side-effect, they can inform patients of the probability of being affected so that patients can make an informed decision about whether to take the drug. To understand the true percentage of people who will be affected by this side-effect, should the researchers conduct a one-sided hypothesis test, or should they conduct a two-sided hypothesis test? Explain.

Solution 2.4.4:

This hypothesis test is related to people’s lives in a most important way. The conclusions made from this hypothesis test can save lives, but potentially at a great cost. It is literally a life and death situation. Therefore, it is extremely important that the conclusions released to the public are backed by the strongest possible evidence. The researchers should select a small significance level, $\alpha$, in order to require a small p-value to achieve statistical significance. Furthermore, using a two-sided hypothesis test would ensure that even stronger evidence (a sample statistic that is even further in the tail of the null distribution) is required to achieve statistical significance.
As of 2017, the average (mean) student loan debt (for current and previous students) in the U.S. was $32,731. We want to know whether the mean is higher now. Suppose we collect a random sample of 150 current and previous students, and our sample has a mean of $34,652.80 of student loan debt. A distribution of simulated sample statistics is shown below. The distribution has a standard deviation of 1712.

Answer the following questions together in groups and be ready to share with the class.

(a) Write out the null and alternative hypotheses in both words and symbols.

(b) Calculate the standardized statistic for the observed sample statistic, and include the correct symbol. Is the standard score positive, negative, or 0? Does its value indicate that the observed statistic is unusual/unexpected or does the standard score indicate that the observed statistic is usual/expected? Explain.

(c) Would the observed statistic have a small p-value, or would it have a large p-value? Explain how you know.

(d) Write the conclusion that we should make about the research conjecture (in context).

(e) Suppose Chase now collects a new random sample of 150 current and former students, and his sample also has an average of $34,652.80 of student debt (just like our original sample). However, Chase decides to phrase his research question as follows: “Is the average (mean) student loan debt in the U.S. different than it was in 2017?” Without doing any calculations, in comparison to our original sample statistic, would Chase have stronger or weaker evidence against the null hypothesis? Explain.

(f) Suppose Aiden now collects a new random sample of 150 current and former students, and his sample has an average of $36,226.90 of student debt. Without doing any calculations, in comparison to our original sample statistic, would Aiden have stronger or weaker evidence against the null hypothesis? Explain.

(g) Suppose Chris now collects a new random sample of 75 current and former students, and they have an average of $34,652.80 of student loan debt (just like our original sample). Without doing any calculations, in comparison to our original sample statistic, would Chris have stronger or weaker evidence against the null hypothesis? Explain.
Section 2.4 | Strength of Evidence in a Conclusion

Exercises

1) Determine whether the following statements are true or false. If the statement is false, explain why it is false and correct the statement by changing the bolded terms.

(a) Using a simulation-based test, the p-value for a two-sided hypothesis test will be about half as large as the p-value for the corresponding one-sided hypothesis test.

(b) Compared to a one-sided hypothesis-test, a two-sided hypothesis-test provides stronger evidence against the null hypothesis.

2) 
   (a) Compare and contrast conducting a two-sided hypothesis test versus a one-sided hypothesis test.
   (b) Provide an original example of when it would be inappropriate to use a one-sided test. What is the consequence of running a one-sided test?
   (c) Provide an original example of when it would be inappropriate to use a two-sided test. What is the consequence of running a two-sided test?

3) Suppose you are testing the hypothesis $H_0: \pi = 0.5$ versus $H_a: \pi > 0.5$. You calculate a sample proportion of $\hat{p} = 0.54$ and find that the p-value is 0.08. Now suppose you redo your study with each of the following changes. Will the new p-value be larger or smaller than the 0.08 you first obtained? Explain your answers using complete sentences.

   (a) You increase the sample size and still find a sample proportion of 0.54. Your new p-value will be ___________. (larger/smaller)
   (b) Keeping the sample size the same, you take a new sample and find a sample proportion of 0.59 instead of 0.54. Your new p-value will be ___________. (larger/smaller)
   (c) With your original sample, you decide to test a two-sided alternative, $H_a: \pi \neq 0.5$, instead of $H_a: \pi > 0.5$. Your new p-value will be ___________. (larger/smaller)

4) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo’s current marketing data indicates that 2 out of every 5 people prefer their brand of cola. They would like to conduct a study to determine whether they can claim that more than 2 out of 5 people prefer Pepsi. A random sample of 318 people is given the taste test, and 138 of them (43.4%) choose Pepsi as their preferred cola.

   Note that each of the scenarios below is independent of the others and should be compared to the original study. How would the strength of evidence change in each scenario? Justify your answer using the methods and reasoning from this section. Write your answers in complete sentences.

   (a) 150 out of 318 people chose Pepsi over Coca-Cola.
   (b) A sample size of 1272 is used, and 552 participants choose Pepsi over Coke.
   (c) The research question is “Is the proportion of the population who prefer Pepsi over Coke equal to 0.4, or something else?”
5) Beth and Rick are human resource analysts tasked with estimating the average commute time of all employees at Company A. The last survey was conducted 5 years ago and found an average commute time of 30 minutes with a standard deviation of 7.2 minutes. Beth and Rick each conduct a survey using a simple random sample.

Beth uses a sample of 10 employees and finds a mean commute of 27.1 minutes with a sample standard deviation of 9.5 minutes.

Rick uses a sample of 300 employees and obtains a mean commute time of 29.3 minutes with a sample standard deviation of 6.9 minutes.

Which analyst (Beth or Rick) has an observed statistic that provides stronger evidence against the hypothesis “The average commute time of all employees is 30 minutes”? Explain.

6) You want to determine whether your dog, Hairy Pawter, understands verbal directions. You take three cups and hide a single treat under one of them, then call out “left” “middle” or “right” depending on which cup the treat is hidden under. If Hairy selects the correct cup, he gets the treat. In your first round of testing you conduct this experiment 25 times and Hairy chooses the correct cup 14 times. The next day you conduct the experiment 100 times and Hairy chooses the correct cup 56 times.

(a) State the null and alternative hypotheses in words and symbols for testing whether Hairy will select the correct cup more often in the long run than if he was just guessing.

(b) Which day of experiments will provide stronger evidence to reject the null hypothesis in favor of the alternative hypothesis? Why?

(c) How would the strength of evidence change if the alternative hypothesis were “The long-run proportion of times Hairy selects the correct cup is something other than 1/3”? Explain.

7) Kelley Blue Book is a vehicle valuation and automotive research company which reports market value prices for new and used automobiles. These values are often used by auto dealerships to determine the selling prices of their vehicles. The average Kelley Blue Book value of a used 2008-2012 sedan is $5,560. I would like to determine whether I can ask for a higher selling price in my region for my sedan, so I conduct a random sample of local ads for used 2008-2012 sedans and record the asking price for 28 of them, obtaining an average asking price of $5,985.

(a) State the null and alternative hypotheses for this scenario.

(b) If the research question were “I would like to determine whether I need to ask for a different selling price in my region”, how would the null and alternative hypotheses change, if at all? How would this change impact the strength of evidence, if at all?

(c) If I were to redo this study using a sample size of 25 rather than 28, what effect (if any) would that have on the strength of evidence against the null hypothesis if all other values remained the same?
For questions 8 through 12, read the original scenario, then explain how the strength of evidence would change when the alternative scenario is used.

8) **Original Scenario:** A study was conducted by the National Endowment for the Arts to describe music preferences in the Hudson Valley. The 2010 report describes relationships between key demographic characteristics and music preferences. Researchers found that 24% of adults preferred the genre of Rock/Heavy Metal at that time. Recently, a random sample of 180 Hudson Valley adults was surveyed with 56 indicating they prefer Rock/Heavy Metal. You want to determine whether the proportion of Hudson Valley residents that prefer Rock/Heavy Metal has changed since the 2010 report.

   **Alternative Scenario:** A random sample of 180 Hudson Valley adults found that 57 of them prefer Rock/Heavy Metal.

9) **Original Scenario:** The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that people are more likely to prefer their brand of cola to Coca-Cola. A random sample of 351 people is given the taste test, and 54% of them choose Pepsi as their preferred cola.

   **Alternative Scenario:** A random sample of 513 people participate in the challenge, and 54% of them choose Pepsi as their preferred cola.

10) **Original Scenario:** The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years with a standard deviation of 9.4 years. After learning this you decide to collect data from a random sample of New York City obituaries from 2018-2020 to compare the life expectancy of New York City residents to all United States citizens. After surveying 75 obituaries, you find a mean life expectancy of 81.7 years.

   **Alternative Scenario:** The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years with a standard deviation of 13.1 years.

11) **Original Scenario:** A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest was for workers who live in New York City, who traveled an average of 35.9 minutes each day with a standard deviation of 12.5 minutes. You live relatively close to the city and know that many of your neighbors work there, so you decide to investigate whether people in your neighborhood have a longer average commute time than the average worker who lives in New York City. You survey 38 commuters in your neighborhood and find an average commute time of 39.7 minutes, with a standard deviation of 6.1 minutes.

   **Alternative Scenario:** The research question is “Do commuters in my neighborhood have a different commute time than 35.9 minutes?”

12) A 2017 survey conducted by the National Science Foundation claimed that fewer than 15% of those currently employed in a STEM (Science, Technology, Engineering and Mathematics) field in the United States are under the age of 30. You conduct a similar survey to determine whether the proportion of STEM field workers in your state who are under the age of 30 is greater than 15%. After collecting a random sample of 210 STEM field workers in your state, you find that 42 are under the age of 30.

   **Alternative Scenario:** After collecting a random sample of 70 STEM field workers in your state, you find that 14 are under the age of 30.
Chapter 3  Theory-Based Approach and Comparing Two Parameters

Contents of Chapter 3

Section 3.1  Sampling

Section 3.2  Association vs. Causation

Section 3.3  Theory-Based Approach to Inference

Section 3.4  Comparing Two Means

Section 3.5  Comparing Two Proportions
**Section 3.1 | Sampling**

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
</table>
| • Fluently use the following vocabulary:  
  O representative sample  
  O biased sampling method  
  O random sample  
• Sampling  
  O Understand why we use a representative sample.  
  O Recognize/identify a representative sample vs. a sample from a biased sampling method.  
  O Identify whether a biased sampling method is likely to overestimate or underestimate the true value of the population parameter and explain why. |}

**Key Terms**

**Representative Sample:**
A subset of the population that accurately “reflects” the larger population. A representative sample will be a miniature version of the population. The sample tends to “look like” the population if the sample is gathered in a representative manner.
- If we collect a sample randomly, then we will be able to say that our sample is representative of the population.

**Sampling Bias:**
A sampling method is biased if it produces sample statistics that consistently overestimate or consistently underestimate the true value of the population parameter.
- The only sampling method that you can trust to not be biased is a method that utilizes randomness.

**Simple Random Sampling:**
A method of selecting a sample from a population in such a way that every possible sample of a given size that could be selected has an equal chance of being selected.
- **Conclusions we can make when using simple random samples:** When we use a simple random sample, we can feel comfortable generalizing our conclusions to the entire population because we are confident that our sample “looks like” the entire population.
- **Important note about simple random sampling:** We need to avoid convenience sampling or using any type of human judgement to select a sample because this leads to sampling bias even when we try to avoid it. It is best to assign each member of the population a number, and then pick the sample using a random number generator that selects members from the population.
- **Important note about sample size:** It is possible to get a representative sample from the population even if we have a “small” sample size!!! You will see in the Collaborative Exercise at the end of this section that even when we collect a small sample using simple random sampling, we will have sample statistics that are very close to the true population parameter.
Sampling

Gathering information from an entire population is often expensive or impossible. Instead, we use a sample of the population. A sample should have the same characteristics as the population it represents. In other words, we want our sample to “look like” our population in every way that we can think of AND cannot think of. Consider the picture below:

Notice how each color takes up the same proportion (percent) of each circle. In this way, the sample “looks like” the population.

It may seem from this simple coloring example that it would not be hard to guarantee that your sample is representative as we can just force these proportions to work out. While that is true, this coloring system corresponds to only one variable (characteristic) of the population. In real life scenarios, there are many variables that we know can have an effect on our study or experiment, and many more which we don’t know! Trying to hand pick a sample that will be representative for all variables (known and unknown) is impossible!

But have no fear for statistics is here! Statisticians use various methods of random sampling to achieve the goal of collecting a representative sample from the population. There are several different methods of random sampling. In each form of random sampling, each member of a population has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called simple random sampling. Any group of individuals is equally likely to be chosen as any other group of the same size if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.
Example 3.1.1

Suppose Lisa wants to form a four-person study group (herself and three other people) from her statistics class, which has 31 members not including Lisa. Describe two ways that Lisa could take a simple random sample of her classmates in order to create her group.

Solution 3.1.1 a)

To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names.

Solution 3.1.1 b)

A more modern way is for Lisa to first list the last names (assuming no duplicate last names!) of the members of her class together with a two-digit number, as shown in the table.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>ID</th>
<th>Name</th>
<th>ID</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>Anselmo</td>
<td>11</td>
<td>King</td>
<td>21</td>
<td>Roquero</td>
</tr>
<tr>
<td>01</td>
<td>Bautista</td>
<td>12</td>
<td>Legeny</td>
<td>22</td>
<td>Roth</td>
</tr>
<tr>
<td>02</td>
<td>Bayani</td>
<td>13</td>
<td>Lundquist</td>
<td>23</td>
<td>Rowell</td>
</tr>
<tr>
<td>03</td>
<td>Cheng</td>
<td>14</td>
<td>Macierz</td>
<td>24</td>
<td>Salangasang</td>
</tr>
<tr>
<td>04</td>
<td>Cuarismo</td>
<td>15</td>
<td>Motogawa</td>
<td>25</td>
<td>Slade</td>
</tr>
<tr>
<td>05</td>
<td>Cunningham</td>
<td>16</td>
<td>Okimoto</td>
<td>26</td>
<td>Stratcher</td>
</tr>
<tr>
<td>06</td>
<td>Fontecha</td>
<td>17</td>
<td>Patel</td>
<td>27</td>
<td>Tallai</td>
</tr>
<tr>
<td>07</td>
<td>Hong</td>
<td>18</td>
<td>Price</td>
<td>28</td>
<td>Tran</td>
</tr>
<tr>
<td>08</td>
<td>Hoobler</td>
<td>19</td>
<td>Quizon</td>
<td>29</td>
<td>Wai</td>
</tr>
<tr>
<td>09</td>
<td>Jiao</td>
<td>20</td>
<td>Reyes</td>
<td>30</td>
<td>Wood</td>
</tr>
<tr>
<td>10</td>
<td>Khan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lisa can use a random number generator to pick three random numbers from 0 to 30. Each of these numbers will correspond to a name of one of Lisa’s classmates.

For example, we use the random number generator at random.org

In this example, we get the numbers 15, 23, and 1. Thus, Lisa’s group will consist of Motogawa, Rowell, and Bautista.
3.1.1

a) From the table of quiz scores below, select 15 at random (do this yourself, try your best!). Then find the average of your chosen quiz scores.

b) Use the random number generator (random.org) to select 15 quiz scores and find the average of the 15 scores.

Hint: You will need to number each score (see Example 3.1.1) in order to use the random number generator!

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

c) How do your averages from a) and b) compare to the average of ALL the quiz scores, which is 7.98?
Example 3.1.2

Determine whether the following are simple random samples. If they are not, discuss whether statistics calculated from the samples will likely over- or underestimate the value of the parameter.

a) A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

b) A student gives a short algebra quiz to classmates in his calculus class to determine how well students at his school know algebra.

c) A teacher puts the names of every student in their statistics course in a hat and then selects 5 of the names to be in a group together.

d) A teacher puts the names of every student in their statistics courses into a hat and then selects a sample to determine the average number of credits a student at their school takes.

Solution 3.1.2

a) This is an example of simple random sampling. Every group of 50 students has an equal chance to be selected.

b) This is not a simple random sample since the students all come from the same class. By the definition, it must be true that any group of students is equally likely to be chosen as any other group of students. Students in the calculus class will probably do better on average on the algebra quiz than the entire student body, so this sampling method will likely produce an overestimate of the parameter. Some students in the school may not have taken algebra or any math class in many years, and those students are not being represented in this sample.

c) This is a simple random sample. Each group of 5 students is equally likely to be picked for the group.

d) This is not a simple random sample. Each student in the teacher’s statistics courses has an equal chance to be selected, but students who are not in those statistics courses have no chance of being selected. This may lead to an overestimate of the average number of credits taken. Students taking statistics may be in majors for which you need to take more credits or may be more likely to be full-time students.

Try It

3.1.2

Determine whether or not the following are simple random samples. If they are not, you should be able to state the reason(s) and also determine whether this sampling method will likely lead to an over- or underestimate of the value of the parameter of interest.

a) To find the average GPA of all students at a university, use all honor students at the university as the sample.

b) To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.

c) To determine the average cost of a two-day stay in a Massachusetts hospital, survey 100 hospitals across the state by creating a list of all hospitals and randomly selecting 100 of them.
In this exercise, the students in the class will each sample words from the list below and then come together to create a sampling distribution.

(a) Each student should pick out 10 words from the list below to form a representative sample:

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>statisticians</td>
<td>26</td>
</tr>
<tr>
<td>use</td>
<td>52</td>
</tr>
<tr>
<td>various</td>
<td>78</td>
</tr>
<tr>
<td>methods</td>
<td>54</td>
</tr>
<tr>
<td>of</td>
<td>55</td>
</tr>
<tr>
<td>random</td>
<td>46</td>
</tr>
<tr>
<td>achieve</td>
<td>59</td>
</tr>
<tr>
<td>the</td>
<td>60</td>
</tr>
<tr>
<td>goal</td>
<td>63</td>
</tr>
<tr>
<td>of</td>
<td>58</td>
</tr>
<tr>
<td>achieve</td>
<td>61</td>
</tr>
<tr>
<td>a</td>
<td>62</td>
</tr>
<tr>
<td>sampling</td>
<td>63</td>
</tr>
<tr>
<td>population</td>
<td>64</td>
</tr>
<tr>
<td>has</td>
<td>52</td>
</tr>
<tr>
<td>pros</td>
<td>77</td>
</tr>
<tr>
<td>and</td>
<td>55</td>
</tr>
<tr>
<td>group</td>
<td>80</td>
</tr>
<tr>
<td>-</td>
<td>81</td>
</tr>
<tr>
<td>chance</td>
<td>103</td>
</tr>
<tr>
<td>of</td>
<td>79</td>
</tr>
<tr>
<td>-</td>
<td>82</td>
</tr>
<tr>
<td>size</td>
<td>107</td>
</tr>
<tr>
<td>to</td>
<td>58</td>
</tr>
<tr>
<td>if</td>
<td>83</td>
</tr>
<tr>
<td>to</td>
<td>108</td>
</tr>
<tr>
<td>describe</td>
<td>84</td>
</tr>
<tr>
<td>is</td>
<td>85</td>
</tr>
<tr>
<td>simple</td>
<td>110</td>
</tr>
<tr>
<td>pick</td>
<td></td>
</tr>
<tr>
<td>called</td>
<td>61</td>
</tr>
<tr>
<td>random</td>
<td>86</td>
</tr>
<tr>
<td>a</td>
<td>111</td>
</tr>
<tr>
<td>method</td>
<td>87</td>
</tr>
<tr>
<td>technique</td>
<td>113</td>
</tr>
<tr>
<td>sample</td>
<td>88</td>
</tr>
<tr>
<td>of</td>
<td>63</td>
</tr>
<tr>
<td>randomly</td>
<td>112</td>
</tr>
<tr>
<td>a</td>
<td>66</td>
</tr>
<tr>
<td>group</td>
<td>91</td>
</tr>
<tr>
<td>in</td>
<td>116</td>
</tr>
<tr>
<td>representative</td>
<td>65</td>
</tr>
<tr>
<td>any</td>
<td>90</td>
</tr>
<tr>
<td>used</td>
<td>115</td>
</tr>
<tr>
<td>be</td>
<td></td>
</tr>
<tr>
<td>equal</td>
<td>41</td>
</tr>
<tr>
<td>of</td>
<td>67</td>
</tr>
<tr>
<td>other</td>
<td>92</td>
</tr>
<tr>
<td>for</td>
<td></td>
</tr>
<tr>
<td>individuals</td>
<td>68</td>
</tr>
<tr>
<td>words</td>
<td>93</td>
</tr>
<tr>
<td>of</td>
<td>46</td>
</tr>
<tr>
<td>each</td>
<td>94</td>
</tr>
<tr>
<td>variable</td>
<td>119</td>
</tr>
<tr>
<td>selected</td>
<td>45</td>
</tr>
<tr>
<td>equally</td>
<td>70</td>
</tr>
<tr>
<td>sample</td>
<td>95</td>
</tr>
<tr>
<td>both</td>
<td>120</td>
</tr>
<tr>
<td>for</td>
<td>47</td>
</tr>
<tr>
<td>likely</td>
<td>71</td>
</tr>
<tr>
<td>of</td>
<td>96</td>
</tr>
<tr>
<td>known</td>
<td>121</td>
</tr>
<tr>
<td>and</td>
<td>97</td>
</tr>
<tr>
<td>unknown</td>
<td>123</td>
</tr>
<tr>
<td>size</td>
<td>98</td>
</tr>
<tr>
<td>is</td>
<td>124</td>
</tr>
<tr>
<td>methods</td>
<td>49</td>
</tr>
<tr>
<td>chosen</td>
<td>99</td>
</tr>
<tr>
<td>impossible</td>
<td>125</td>
</tr>
</tbody>
</table>

(b) For each word chosen, write down the word as well as the length of the word (number of letters).

(c) Compute the average length of the 10 words that you picked. Include the correct symbol here!

Average Length =

(d) Add your sample statistic (average length) to the class dot plot.

(Continued on next page)
(e) The average length of all the words in the list is 4.82 letters. Is this a parameter or statistic? What is the correct symbol to use for this number?

(f) How does the average 4.82 letters compare with your sample mean word length? How about the mean of the distribution created by your class?

(g) Are your answers to (f) evidence that your method for sampling words was biased? If so, did the sampling method tend to produce statistics that were over- or underestimates for the parameter of interest?

(h) We are now going to sample again, but this time randomly. There are 125 words in the list. Use this fact and the random number generator (random.org) to select 10 new words at random.

<table>
<thead>
<tr>
<th>Random Number</th>
<th>Word</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(i) Compute the average length of the 10 words in your random sample. Include the correct symbol here!

\[
\text{Average Length} = \]

(j) Add your sample statistic (average length) to the new class dot plot.

(k) What is the center of the new class distribution? How does this compare to the population parameter of 4.82 letters? What is the shape of the new class distribution?

(l) Based on the shape and center of the new class distribution, do you believe that simple random sampling is a biased sampling method? Explain.
Section 3.1 | Sampling

Exercises

1)  **(True or False)** Determine whether the following statements are true or false. If the statement is false, explain why.
   (a) A large convenience sample is just as likely to represent the population as a small random sample.
   (b) A sample is always a subset of the population.
   (c) If you do not use randomness when collecting your sample, it will not be a representative sample.
   (d) Increasing the sample size of a sample will decrease the bias in that sampling method.

2)  Sara would like to collect a sample for her study. Her colleague, Gary, suggests that she use a simple random sample, and Sara agrees. Which of the following methods could Sara use to help ensure she is selecting a simple random sample? If it will not create a simple random sample, explain why not.
   (a) Choosing every 20th name from a database
   (b) Looking at subjects’ profiles online and choosing a variety of participants
   (c) Flipping a coin for each entry in a database and selecting them if the coin lands on “heads”.
   (d) Closing her eyes and pointing at names from a list of DCC students

3)  To determine how many minutes per day SUNY students exercise on average, an instructor has all the students in his statistics class (22 students) take a short survey. The instructor finds that students in his statistics class reported exercising 19 minutes per day on average.
   (a) Was this a simple random sample? Explain.
   (b) Would you expect this statistic to be an overestimate, underestimate, or an accurate estimate of the population proportion? Explain.
   (c) What population(s) (if any) could the results of this study be applied to?

4)  Take a moment to define what it means to for a sampling method to be biased. In each scenario below, identify (describe in words) the population of interest, determine whether the sampling method will likely create a representative sample, and determine whether the statistic will likely overestimate, underestimate, or accurately estimate the parameter of interest. Explain your reasoning.
   (a) To determine the proportion of DCC students who have a primary residence outside of Dutchess County you collect a sample of 80 students as they exit the residence hall.
   (b) While investigating the average age of customers at a local fast food restaurant you survey people as they enter the drive-through.
   (c) You record the method of arrival at an emergency room entrance for every 3rd person every day for a week.
   (d) To determine what the preferred ice cream flavor is in your town, you wait outside an ice cream parlor and flip a coin every time someone exits. If the coin lands on heads, you ask them their favorite ice cream flavor and record the response until you get 40 responses.
   (e) To determine the average amount of time DCC students study each day we use a random number generator and a database to select and survey 45 first-year students during a meeting with their advisor.
   (f) To determine the average height of all students in a statistics class, the heights of the first five people to walk into the room are recorded.
5) Suppose you want to determine the proportion of all Poughkeepsie residents ages 18 and older who work more than one job. Explain how you would conduct your survey to obtain a representative sample of the population. Be sure to explain how you would collect your sample. Where do you go? When? Who do you select and why?

6) In 1948 the Chicago Daily Tribune erroneously published a headline due to early polling results that indicated Thomas E. Dewey had defeated Harry S. Truman in that year’s presidential election. In truth, Truman won 49.6% of the popular vote (24,179,347 out of 47,346,569 votes) to Dewey’s 45.1% (21,991,292 out of 47,346,569 votes) with the remaining percentage going to third parties.

The Gallup’s September 24 report indicated that they had obtained a random sample of 3,250 voters, and 46.5% of sampled voters indicated they would be voting for Dewey, while only 38% planned to vote for Truman.

(a) The statistic for the proportion of voters who would select Dewey was nearly correct, but Truman’s proportion was off by over 11 percentage points. Does this mean the sample was not representative of the population of U.S. voters on September 24, 1948? What are some possible explanations for why this survey did not accurately predict the results of the election?

(b) Let us assume that Gallup did collect a representative sample of U.S. voters. What could have happened between September 24, 1948 and November 2, 1948 (election day) that could explain the discrepancy between the survey and the results?

7) A group of National Fish and Wildlife researchers are collecting data on the bear population in Bearenstein National Park. It is estimated that there are 640 grizzly bears in the park. Recently, the researchers have noticed a parasite among several of the bears and need to determine how widespread the parasite is among all the bears. The researchers found that of 29 bears surveyed, 16 were male and 13 were female, the average weight of the bears was 692.5 pounds, and 5 of the 29 bears have the parasite.

(a) Identify (describe in words) the population the researchers are interested in.

(b) Identify (describe in words) the parameter and statistic of interest and label them using the appropriate symbols.

(c) The researchers want to determine whether there is evidence that fewer than one-fifth of the bear population has the parasite. Determine the null and alternative hypotheses for this test. Write the hypotheses using both words and symbols.

(d) Use the One Proportion applet to conduct a statistical analysis to answer the research question. State the p-value obtained and use it to answer the research question.

(e) To what population, if any, are you comfortable extending these results?
Key Terms

**Association**: Two variables are associated if knowing the value of one variable can tell you about the value of the other variable. Associated variables are “connected” and when displayed graphically, seem to follow a pattern together.

**Causation**: Two variables have a causal relationship if they are associated and a change in one variable causes a change in the other variable.

**Response Variable**: The variable we believe is being impacted or changed by the explanatory variable.

**Explanatory Variable**: The variable we believe is “explaining” the change in the response variable.

**Confounding Variable**: A variable that is related to both the explanatory variable and the response variable in such a way that its effects on the response variable cannot be separated from the effects on the explanatory variable.

**Observational Study**: A study in which the values of the explanatory variable are simply observed.

**Experiment**: A study in which the researchers assign subjects to different treatment groups based on the explanatory variable.

**Random Assignment**: An experimental design in which the researcher randomly assigns each subject to a treatment group.
Association and Causation

In general, if we see a pattern between two variables, what can we say about the cause of the pattern? In statistics, and science in general, we are often searching for a cause-and-effect relationship.

Consider the following chart:

![Chart showing the correlation between coffee consumption and deaths from misusing nonpowered hand tools over the years from 2000 to 2009.](http://tylervigen.com/view_correlation?id=56904)

Source: [http://tylervigen.com/view_correlation?id=56904](http://tylervigen.com/view_correlation?id=56904)

From 2000 to 2009, there is a clear pattern between the amount of coffee an average American consumes and the number of people killed by misusing a nonpowered hand tool.

This is an example of an association between variables. Starting from 2000, if we trace the red line, the yellow line can be thought of as “following along” (at least until 2005). Thus, knowing the value of the red curve for a particular year can give us information about the value of the yellow curve for that same year.

But is this a cause-and-effect relationship? Was there a drop in coffee consumption in 2001 because fewer people were killed by misusing a nonpowered hand tool? This seems very unlikely! This is an example of association and not causation. Variables can be connected/associated, but that connection/association does not have to come from a cause-and-effect relationship!

So, what is happening here?

In situations like this we are often looking for a confounding variable. One way to think of the confounding variable is as an alternative explanation for the association we see between the explanatory and response variables.
Explanations 1. and 2. describe a cause-and-effect relationship between A and B. The third explanation involves a confounding variable, C, that is causing the association between A and B.
Example 3.2.1

There is an association between the ice cream consumption in New York City and the murder rate. To be more specific, as the consumption of ice cream rises so do the number of murders. Is it reasonable to believe that one of these variables is causing the change in the other? If so, one of these two sentences would sound believable:

- The more ice cream you eat, the more likely you are to commit murder.
- Those who commit murder eat more ice cream than those who don’t commit murder.

If we do not believe either of these explanations, can we find a confounding variable at play here? In other words, could there be another variable that is creating the association that we see between the murder rate and the rate of ice cream consumption?

Solution 3.2.1

One possible confounding variable is temperature. Both of these increases (ice cream consumption and murder) happen during the summer months. What do we need to check in order to conclude that temperature is a confounding variable?

a) The temperature rising will cause an increase in the ice cream consumption.
b) The temperature rising will cause an increase in the murder rate.
c) The changes in 1. and 2. must agree with the association we see in our two variables.

For a), an increase in temperature will cause an increase in ice cream consumption.

For b), an increase in temperature will cause an increase in murder rate. See the linked source for some possible explanations why.

For c), as the temperature rises, by a) and b), both ice cream consumption and murder rate will increase together. A rise in the ice cream consumption will correspond with a rise in the murder rate, but only because a rise in the temperature is causing both!

The situation in example 3.2.1 is an observational study. Those who gathered the data and reported on it had no control over the variables. They simply observed. In an observational study we are not able to determine a cause-and-effect relationship between two variables. Since we do not have control over the observational units, we have no way to separate any possible confounding variables from the situation, and so it may be the case that a confounding variable is creating the association that we observe.
Experiments

To explore a relationship without worrying about the impact of possible confounding variables, researchers will conduct experiments. The purpose of an experiment is to investigate a relationship between two variables. When changes in one variable cause changes in another variable, we call the first variable the explanatory variable. The affected variable is called the response variable. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different outcomes or values of the explanatory variable are called treatments.

Even if the association between two variables is not causal, we can still designate one variable as explanatory and the other variable as response.

Example 3.2.2

In the following scenarios, identify the explanatory and response variables and classify each variable as categorical or quantitative.

a) Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers calculate the proportion men in each group who have had heart attacks.

b) The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed pencil and paper mazes multiple times while wearing masks. They completed the mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject’s impression of the mask’s scent: positive, negative, or neutral.

Solution 3.2.2

a) In this scenario the explanatory variable is the medication given (a categorical variable with outcomes aspirin and placebo). The response variable is whether the participant had a heart attack (also a categorical variable). Researchers are interested in whether taking aspirin can lower your chances of having a heart attack.

b) In this scenario the explanatory variable is mask scent (a categorical variable with outcomes floral and unscented). The response variable is the time it takes to complete the maze (a quantitative variable). Researchers are interested in whether having a particular kind of scent can affect the time it takes to complete a pencil and paper maze.
How an Experiment with Random Assignment Leads to Ability to Conclude Causation

In order to have the ability to claim a causal relationship, researchers must remove all of the possible confounding variables. If we can’t even think of all the possible confounding variables that could be causing the association, then how can we possibly account for them all?

Let’s first start with an example of what could happen without random assignment. Consider a study in which we want to test the effectiveness of a new drug at preventing symptoms for a certain disease. The researchers gather subjects and place them into two treatment groups without using random assignment. It turns out that 90% of the group who receives the drug is over 5’7” while only 30% of the placebo group is over 5’7”. The study is finished, and the researchers find a statistically significant improvement in the group assigned the new drug.

Can we conclude that the improvement was caused by the drug? If the only difference between these groups is that one received the drug and the other a placebo, then the answer is yes. In this case we could have many differences. We know that the drug group had a much higher percentage of people over 5’7”. Could it be that being taller makes you less likely to have symptoms? The drug group may also be heavier on average than the placebo group or contain a higher proportion of men than women. Each of these could be the cause for the better results in the drug group. If there are other variables that are potentially imbalanced between the groups, we cannot claim that the drug is causing the improvement!

For the example above, it may seem like we can fix this by having the researchers rearrange the groups until the heights are balanced between them. At this point, we would be comfortable knowing that this variable will not be the cause of the improvement. The researchers look over the groups and realize that they have imbalanced the groups with respect to age! The drug group is much younger on average than the placebo group. If researchers run the experiment with these groups, we won’t know if it is the youth of the group or the drug that is causing the improvement. For every variable we could think of, such gender, medical history, height, weight, or age, we would need to attempt to balance each variable across the two treatment groups in order to be sure that the drug caused the improvement. We also must somehow account for all of the other variables that we haven’t thought of that could impact improvement!

So, what can we do? The ‘magic’ here is all in the process of random assignment. If you randomly assign the subjects in your experiment to treatment groups, you will ‘balance’ (not always perfectly!) the groups on every variable except for the one you control (explanatory variable). As an example, let’s say we are running a study with two treatment groups: Treatment Group 1 and Treatment Group 2. Suppose we have 20 participants: 10 men and 10 women. We then randomly assign these 20 participants to the two treatment groups three times:

The first time, we create the two treatment groups as follows:
(Treatment group 1 has 4/10 = 40% female, and 6/10 = 60% male)

<table>
<thead>
<tr>
<th>Group 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 M 2 F 3 F 4 F 5 M 6 M 7 M 8 M 9 F 10 M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 M 2 F 3 M 4 M 5 F 6 F 7 F 8 M 9 F 10 F</td>
</tr>
</tbody>
</table>
The second time, we create the two treatment groups as follows:
(Treatment group 1 has 5/10 = 50% female, and 5/10 = 50% male)

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 F</td>
<td>1 F</td>
</tr>
<tr>
<td>2 F</td>
<td>2 F</td>
</tr>
<tr>
<td>3 M</td>
<td>3 F</td>
</tr>
<tr>
<td>4 F</td>
<td>4 M</td>
</tr>
<tr>
<td>5 M</td>
<td>5 F</td>
</tr>
<tr>
<td>6 F</td>
<td>6 M</td>
</tr>
<tr>
<td>7 M</td>
<td>7 F</td>
</tr>
<tr>
<td>8 M</td>
<td>8 F</td>
</tr>
<tr>
<td>9 F</td>
<td>9 M</td>
</tr>
<tr>
<td>10 M</td>
<td>10 M</td>
</tr>
</tbody>
</table>

The third time, we create the two treatment groups as follows:
(Treatment group 1 has 6/10 = 60% female, and 4/10 = 40% male)

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 M</td>
<td>1 M</td>
</tr>
<tr>
<td>2 F</td>
<td>2 F</td>
</tr>
<tr>
<td>3 M</td>
<td>3 M</td>
</tr>
<tr>
<td>4 M</td>
<td>4 F</td>
</tr>
<tr>
<td>5 F</td>
<td>5 M</td>
</tr>
<tr>
<td>6 F</td>
<td>6 F</td>
</tr>
<tr>
<td>7 F</td>
<td>7 F</td>
</tr>
<tr>
<td>8 F</td>
<td>8 M</td>
</tr>
<tr>
<td>9 F</td>
<td>9 M</td>
</tr>
<tr>
<td>10 M</td>
<td>10 M</td>
</tr>
</tbody>
</table>

**Randomization 1:**
Group 1: 60% M, 40% F  
Group 2: 40% M, 60% F

**Randomization 2:**
Group 1: 50% M, 50% F  
Group 2: 50% M, 50% F

**Randomization 3:**
Group 1: 40% M, 60% F  
Group 2: 60% M, 40% F

What we notice is that while not perfect, creating the treatment groups using random assignment approximately balances sex across the two groups.
Causation

If we do run a randomized experiment (i.e., an experiment where the explanatory variable groups are formed using random assignment), and we find that there is a statistically significant association between the explanatory and response variables, then we can claim a cause-and-effect relationship between the explanatory and response variables. If we randomly assign subjects to treatment groups, then the only difference between the groups will be the difference in the explanatory variable categories. Any statistically significant association can then be attributed to the difference in the explanatory variable and not to any confounding variable!

Example 3.2.3

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution 3.2.3

The explanatory variable is birth order. You cannot randomly assign a person’s birth order. Only random assignment eliminates the impact of confounding variables. When you cannot assign subjects to treatment groups at random, then there may be a confounding variable that is causing the association between the explanatory and response variables.
You are concerned about the effects of texting on driving performance. You want to understand the following: “How many seconds does it take for a driver in a following car to respond, on average, when a leading car hits the brakes?” Answer the questions below in order to design a study to test the response time of drivers while texting vs. while driving without texting.

a) Describe the explanatory and response variables in the study.

b) Suppose we created one experimental group that was all blondes, and another experimental group that was all non-blondes. Suppose we then had the blondes text while driving and measured their response times, and had the non-blondes drive without texting and measured their response times. If we found that the non-texting group had a shorter response time, could we conclude that not texting causes a shorter response time? Explain.

Be Careful! Common Error Alert!

A common error is to confuse Random Sampling and Random Assignment

- If we use random sampling, then we can be confident that we have selected a representative sample that “looks like” the larger population, and so we can use the sample to make conclusions about the entire population.

- If we use random assignment to create the experimental group(s), and we conclude that a statistically significant association (correlation) exists, then we can also conclude that the explanatory variable causes the effect we see in the response variable (rather than only concluding association).

Random Sampling → We can use our sample to make conclusions about the population.

Random Assignment → We can conclude causation instead of only concluding association.
Section 3.2 | Association vs Causation

“Correlation doesn’t imply causation, but it does waggle its eyebrows and gesture furtively while mouthing ‘look over there’.”
- Randall Munroe

Exercises

1) Each of the headlines below were presented as causation by newspapers, magazines, news stations, or websites. In most cases, cause and effect between two variables was improperly assumed or the results were extended to a population to which they do not apply. For each headline complete parts (a) through (d).

(a) Identify the observational units of the study.
(b) Identify the explanatory variable and classify it as categorical or quantitative.
(c) Identify the response variable and classify it as categorical or quantitative.
(d) Provide an example of a potential confounding variable (assuming random assignment was not used).

Recall: A confounding variable is a variable that is associated with both the explanatory variable and the response variable in such a way that its effects on the response variable cannot be separated from its effects on the explanatory variable.

(1.1) Candy Cigarettes Encourage Young People to Smoke
(1.2) Diet of Fish Helps Prevent Teen Violence
(1.3) Sugar Fuels Growth of Cancer
(1.4) Late-Night Snacks Destroy the Part of Your Brain That Stores Memories
(1.5) Hugging Your Dog Increases Stress Levels in Your Dog
(1.6) Glass of Red Wine Equivalent to Hour of Gym Time
(1.7) Scientists Say Smelling Farts Can Help Prevent Cancer
(1.8) Coffee Consumption – in either partner – Increases Miscarriage Risk

2) In May 2014 BBC News reported the headline “Sleeping in a room with too much light has been linked to an increased risk of piling on the pounds”.

(a) Identify the observational units, explanatory variable, and response variable based on the BBC News headline above.

The actual study ([https://academic.oup.com/aje/article/180/3/245/2739112](https://academic.oup.com/aje/article/180/3/245/2739112)) was carried out by the University of Oxford with 113,000 women aged 16 years or older, living in the UK between 2003 and 2012. Researchers asked participants to rate the amount of light in their bedrooms at night as “light enough to read”, “light enough to see across the room but not read”, “light enough to see your hand in front of you but not across the room”, and “too dark to see your hand (or you wear a mask)”. They also collected data on BMI (and a number of other potential confounding variables).

(b) Identify the observational units in the University of Oxford study.
(c) Identify the explanatory variable and classify it as categorical or quantitative.
(d) Identify the response variable and classify it as categorical or quantitative.
(e) “Night shift work in the previous 10 years” was one potentially confounding variable in this study. Explain what it means for this to be a confounding variable in this study, and describe how this could provide an alternative explanation to concluding that that “sleeping in a room with too much light causes weight gain”.
(f) What are some potential confounding variables you would try to account for if you were the researcher?
3) A popular comic by Randall Munroe depicts the following situation.

While sitting in the library you hear the following conversation between your good friends, Gary and Josh:

Gary: I used to think correlation implied causation.
[Gary lifts his hand while continuing to talk to Josh.]
Gary: Then I took a statistics class. Now I don't.
Josh: Sounds like the class helped.
Gary: Well, maybe.

Explain the joke here.

4) Complete parts (a) and (b) for each of the scenarios listed below.

(a) Determine whether the study is likely to be an observational study or an experiment. Justify your answer using complete sentences and the methods/reasoning from this section.

(b) Identify the explanatory and response variables and classify each variable as categorical or quantitative.

(4.1) The effects of local hyperthermia on the blood flow, oxygen pressure, and pH in tissues were investigated using tumor bearing rats.

(4.2) Researchers examined Social Security records to see whether women with the names Virginia, Georgia, Louise, and Florence were especially likely to have moved to the state of Virginia, Georgia, Louisiana, and Florida, respectively.

(4.3) An educational researcher compares the academic performance of students from the two different high schools in a city.

(4.4) A neuroscientist compares people’s ability to recall facts after they either listened to a story or read a story.

(4.5) An instructor compares exam results between their morning class and their afternoon class.

(4.6) A food scientist studies the relationship between the temperature inside a refrigerator and the number of bacteria on the food.

(4.7) An airline varies the amount of leg space available on four otherwise identical planes and surveys passengers on how comfortable their flight was.

(4.8) A social psychologist sat on the sidewalk pretending to be hurt for a day in New York City, Miami, Akron, Kansas City, Denver, and Los Angeles and recorded the proportion of people who stopped to ask if they needed help.

5) In each of the following studies identify the explanatory and response variables and classify each variable as categorical or quantitative. Determine whether the researchers can conclude cause and effect and justify your answers using complete sentences.

(a) In a random sample of countries, it was found that countries where more citizens have Internet access also have higher life expectancies.

(b) Students were randomly assigned to two groups. One group’s exams were printed on blue paper and the other group’s exams were printed on green paper. The group using blue paper scored higher on average than the group using green paper.

(c) An instructor gave one their classes a test while wearing a duck costume and gave another class their test while dressed in a business casual outfit. The class who witnessed the duck costume scored significantly higher on the test than the class who witnessed the business casual outfit.

(d) A survey of social media usage among students found that students with fewer social media accounts had, on average, higher grade point averages.
6) For many years, Big Tobacco (a name used to refer to the largest global tobacco industry companies) funded research to argue that smoking does not cause cancer, but rather, there is a gene which both makes people want to smoke and causes higher rates of cancer in those who have it. Explain how you could design and run an experiment to test this theory.

**Hint:** This is difficult – there is a reason Big Tobacco got away with their actions for years after it was clear smoking causes health issues.
Section 3.3 | Theory-Based Approach to Inference

### Section Objectives

- Identify the conditions necessary for using a theory-based approach to inference.
- Apply the theory-based approach to a single categorical variable.
- Apply the theory-based approach to a single quantitative variable.

---

**Theory Based Approach**

In this section we will discuss a theory-based approach to hypothesis testing that eliminates the need to run simulations. Before we introduce this idea, let’s review why we have been using simulation.

To run a hypothesis test, up to this point we have done the following:

- Create a null hypothesis and an alternative hypothesis.
- Have a computer simulate drawing samples many times under the assumption that the null hypothesis is true in order to create a distribution of simulated sample statistics.
- Compare the observed statistic to the distribution of simulated sample statistics and use a p-value and/or a standardized statistic to determine whether the observed statistic is usual/expected or unusual/unexpected.
- Conclude that the null hypothesis is plausible (observed statistic is usual/expected under the null hypothesis assumption) or conclude that the null hypothesis is not plausible (observed statistic is unusual/unexpected under the null hypothesis assumption).

In the theory-based approach we have **one major difference:**

- Create a null hypothesis and an alternative hypothesis.
- **Create a theoretical distribution of sample statistics based on the assumption that the null hypothesis is true.**
  - Compare the observed statistic to the distribution of simulated sample statistics and use a p-value and/or a standardized statistic to determine whether the observed statistic is usual/expected or unusual/unexpected.
  - Conclude that the null hypothesis is plausible (observed statistic is usual/expected under the null hypothesis assumption) or conclude that the null hypothesis is not plausible (observed statistic is unusual/unexpected under the null hypothesis assumption).

What is this “theoretical sampling distribution” and how is it created? As we have discussed in previous sections, we continue to see a very similar picture from all our simulated chance models.

---

Section 1.2

$H_0: \pi = 0.4$

Section 1.4

$H_0: \pi = 0.8$

Section 1.5

$H_0: \mu = 315$
What we notice from this is that our chance model is approximately bell-shaped and always centered at about the null hypothesis value. This is no coincidence, as the following theorem states.

**Central Limit Theorem**

If random samples of size $n$ are drawn from a population, the distribution of sample statistics approaches a normal distribution as $n$, the sample size, increases. This distribution will have the population parameter as its center (mean).

### Example 3.3.1

In this example, we are investigating whether Mr. Young has more than 80% support from all registered voters. We collect a random sample of registered voters. Our hypotheses are as follows:

- $H_0: \pi = 0.8 \quad$ 80% of all voters will vote for Mr. Young
- $H_A: \pi > 0.8 \quad$ More than 80% of all voters will vote for Mr. Young

(a) Create a distribution of simulated sample statistics with 1000 repetitions if the sample size was 5 registered voters.
(b) Create a distribution of simulated sample statistics with 1000 repetitions if the sample size was 40 registered voters.
(c) Create a distribution of simulated sample statistics with 1000 repetitions if the sample size was 200 registered voters.
(d) Create a distribution of simulated sample statistics with 1000 repetitions if the sample size was 10,000 registered voters.
(e) Compare the distributions from (a) – (d).

### Solution 3.3.1

(a) For a sample of $n = 5$ registered voters with 1000 simulated sample statistics, we see that the distribution of sample statistics does not closely resemble a normal distribution.

(b) For a sample of $n = 40$ registered voters with 1000 simulated sample statistics, we see that the distribution of sample statistics looks much more like a normal distribution.
(Example 3.3.1 continued)

(c) For a sample of $n = 200$ registered voters with 1000 simulated sample statistics, we see that the distribution of sample statistics looks even more like a normal distribution:

(d) For a sample of $n = 10,000$ registered voters with 1000 simulated sample statistics, we see that the distribution of sample statistic very much looks like a normal distribution:

(e) As we increase the sample size, we can see the distribution of sample statistics becoming more bell shaped! This shows us that the distribution of sample statistics will more closely approximate the theoretical normal distribution as the sample size increases. We also notice that the center of each distribution is about 0.80 or 80%. What changes in each distribution is the variability, or standard error.

**Theory Based Approach – One Proportion**

We can create approximations of these distributions using mathematical, theoretical techniques and use these theoretical distributions to do determine whether the observed statistic is usual or unusual. Consider the following facts about theory-based distributions of sample statistics:

- The mean of the theoretical normal distribution is the assumed value of $\pi$ from $H_0$ (just like with the simulation).
- The standard deviation of the theoretical normal distribution (called the standard error) is $\sqrt{\frac{\pi(1-\pi)}{n}}$, where $\pi$ is the assumed value of the parameter from the null hypothesis.
- We can use the p-value or the z-score to make conclusions about the hypotheses.
- The theory-based approach and simulation-based approach will produce basically the same results if certain conditions are met (10 successes and 10 failures in the sample for categorical data). When comparing the simulation-based approach and the theoretical approach, you should see similar sampling distributions, with similar conclusions when the conditions are met.
- We can use the Theory-Based Inference applet to create the normal distribution and find the standardized statistic (z-score) and p-value.
In the previous example, we investigated whether Mr. Young had more than 80% support from all registered voters. We then collected a random sample of 114 registered voters and found that 100 support Mr. Young. Our hypotheses were as follows:

\[ H_0: \pi = 0.8 \quad \text{80% of all voters will vote for Mr. Young} \]
\[ H_A: \pi > 0.8 \quad \text{More than 80% of all voters will vote for Mr. Young} \]

(a) Create a theoretical sampling distribution from this information using the facts preceding this example.
(b) Create a theoretical sampling distribution from this information using the One Proportion applet.
(c) Find the z-score and p-value and make a conclusion in context.

**Solution 3.3.2**

(a) We will first create the theoretical sampling distribution by hand. We know that the distribution will be bell-shaped and the center (mean) of this distribution is assumed in the null hypothesis. We can find the standard deviation (standard error) using the formula given previously:

\[
\sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.8(1-0.8)}{114}} = 0.037
\]

Let’s fill out a normal distribution with this information.

(b) Using the One Proportion applet, we can select the “Normal Approximation” option to see the theoretical sampling distribution.

<table>
<thead>
<tr>
<th>Probability of success (π):</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (n):</td>
<td>114</td>
</tr>
<tr>
<td>Number of samples:</td>
<td>1</td>
</tr>
<tr>
<td><strong>Options:</strong></td>
<td></td>
</tr>
<tr>
<td>- Two-sided</td>
<td></td>
</tr>
<tr>
<td>- Exact Binomial</td>
<td></td>
</tr>
<tr>
<td>- Normal Approximation</td>
<td></td>
</tr>
</tbody>
</table>
(Example 3.3.2 continued)

(c) Let’s find the z-score by hand first.

\[ z = \frac{0.877 - 0.80}{0.037} = 2.081 \]

Since the z-score is greater than 2, the null hypothesis is not plausible and we have strong evidence that more than 80% of registered voters support Mr. Young.

We can also use the Theory-Based Inference applet.

We see here that we get a z-score of 2.06 and a p-value of 0.0197. Both of these tell us we have strong evidence that more than 80% of registered voters support Mr. Young!

**You will notice a slight difference between the z-score you calculated by hand and the z-score from the applet. This difference is due to rounding we did in the by-hand calculation that the computer does not do.**
Theory Based Approach – One Mean

We can do very similar work when considering a single mean. In this case we will be using a theoretical t-distribution to model the distribution of simulated sample statistics (distribution of \( \bar{x} \) values). In order to create and use this distribution we need the following facts:

- The theoretical t-distribution is bell-shaped and the center (mean) of the distribution is the assumed value from the null hypothesis, \( \mu \).

- The standard deviation of the theoretical t-distribution (called the standard error) is \( \frac{s}{\sqrt{n}} \), where \( s \) is the observed sample standard deviation.
  
  Note that we will always have a known sample standard deviation, \( s \), and we rarely have the true population standard deviation, \( \sigma \). So, we are using the standard deviation of the sample to calculate the standard error of the t-distribution. However, if we did have the value of the true population standard deviation, \( \sigma \), then we could use the formula \( \frac{\sigma}{\sqrt{n}} \), where \( \sigma \) is the standard deviation of the population, and the distribution of sample means would be a normal distribution instead of a t-distribution. But, note that it would be a rare situation in real life that we didn’t know the true value of the population mean, \( \mu \), but we did know the true value of the population standard deviation, \( \sigma \)!

- The theory-based approach and simulation-based approach will produce basically the same results if certain conditions are met (sample size of 20 or more and no strong skew in the sample data). When comparing the simulation-based approach and the theoretical approach, you should see similar sampling distributions, with similar conclusions when the conditions are met.

- We can use the Theory-based Inference applet to create the t-distribution and find the standardized statistic (t-score) and p-value.

When working with quantitative variables we are using a new distribution called a t-distribution instead of using the normal distribution. For more information on this, see here¹.

¹ [https://www.geeksforgeeks.org/students-t-distribution-in-statistics/](https://www.geeksforgeeks.org/students-t-distribution-in-statistics/)
A production line is supposed to produce boxes of cereal with an average of 20 ounces of cereal in the boxes. The line supervisor wants to know whether the boxes are coming out with an average of 20 ounces like they are supposed to. The hypotheses are as follows:

\[ H_0: \mu = 20 \quad \text{All the boxes have an average (mean) of 20 ounces of cereal.} \]
\[ H_A: \mu \neq 20 \quad \text{The average (mean) amount of cereal in all boxes is not 20 ounces.} \]

We take a sample of size 40 and find a sample mean of 20.4 ounces and a sample standard deviation of 0.5 ounces.

(a) Create a theoretical sample distribution from this information using the facts preceding this example.

(b) Create a theoretical sample distribution from this information using the Theory-based Inference applet.

(c) Find the t-score and p-value and make a conclusion in context.

**Solution 3.3.3**

(a) We will first create the sampling distribution by hand. We know that the distribution will be bell-shaped and the center (mean) of this distribution is assumed in the null hypothesis. We can find the standard deviation (standard error) using the formula given previously:

\[ s = \frac{s}{\sqrt{n}} = \frac{0.5}{\sqrt{40}} = 0.079 \]

Let’s fill out a t-distribution with this information.

(continued on next page)
(Example 3.3.3 continued)

(b) Using the Theory-based Inference applet,

(c) We can see the t-score, $t = 5.06$, from the applet. Since the t-score is greater than 2, the null hypothesis is not plausible and we have very strong evidence that the average (mean) amount of cereal in all boxes is not 20 ounces. We can also calculate the standardized statistic by hand:

$$ t = \frac{20.4 - 20}{0.079} = 5.06 $$

Notice that the p-value of 0 matches the conclusion that the null hypothesis is not plausible.

You might notice in the screenshot above an extra piece of information, $df = 39$. This refers to the degrees of freedom. We will not spend time focusing on this idea, but it is an important factor in the shape of the t-distribution and can be used to find the t-score from a table.

Try It

3.3.1 Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample of cans are measured and the statistics are $n = 34$, $\bar{x} = 16.01$ ounces $s = 0.143$ ounces.

(a) Create a theoretical sampling distribution from this information using the facts preceding Example 3.3.3.

(b) Create a theoretical sampling distribution from this information using the Theory-based Inference applet.

(c) Find the t-score and p-value and make a conclusion in context.
Caution – Validity Conditions

When a sample of quantitative data satisfies the conditions

\textit{sample size is 20 or more and sample data not strongly skewed}

we can use a THEORETICAL t-distribution to model the distribution of sample statistics (distribution of $\bar{x}$ values). Note that a t-distribution has the same shape as a normal distribution; both are bell-shaped and symmetric.

If these conditions are not met, we should not use the theory-based approach and should instead use simulation.
Section 3.3 | Theory-Based Approach to Inference

Exercises

1) **(True or False)** Determine whether the following statement is true or false. If false, explain why.
When you are using the theory-based approach for your statistical analysis, if you were to switch from a two-sided to a one-sided hypothesis test, your p-value would be exactly twice as large.

2) Andrew and Jessi are testing the hypotheses $H_0: \mu = 3.5$ and $H_A: \mu > 3.5$. They take a sample of size 82 and obtain a sample mean of 3.9 and a sample standard deviation of 0.83. Andrew wants to use an applet to run a simulation for their analysis. Jessi says it is appropriate for them to use the Theory-based Inference applet for their analysis. Explain why either approach should be fine to use in this situation.

3) Imagine you were testing the hypotheses $H_0: \pi = 0.15$ and $H_A: \pi < 0.15$.
(a) State the criteria used to determine whether the theoretical approach would be valid. Be as specific as possible.
(b) Under what conditions would you expect the simulation approach to produce nearly identical results to the theoretical approach? Be as specific as possible.

4) Imagine you were testing the hypotheses $H_0: \pi = 0.42$ and $H_A: \pi \neq 0.42$.
(a) State the criteria used to determine whether the theoretical approach would be valid. Be as specific as possible.
(b) Under what conditions would you expect the simulation approach to produce nearly identical results to the theoretical approach? Be as specific as possible.

5) Imagine you were testing the hypotheses $H_0: \mu = 25.82$ lbs. and $H_A: \mu \neq 25.82$ lbs.
(a) State the criteria used to determine whether the theoretical approach would be valid. Be as specific as possible.
(b) Under what conditions would you expect the simulation approach to produce nearly identical results to the theoretical approach? Be as specific as possible.

6) The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years with a standard deviation of 6.5 years. After learning this you decide to collect data from a random sample of Dutchess County obituaries from 2018-2020 to compare the life expectancy of Dutchess County residents to all United States citizens to determine whether Dutchess County residents tend to live longer or shorter lives compared to the national average. A random sample of 51 local obituaries found an average age at the time of death of 75.6 years with a standard deviation of 10.3 years. The sample data were not strongly skewed.
(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.
(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.
(c) Use the appropriate applet to generate a distribution of sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.
(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting.
in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.

7) A 2017 survey conducted by the National Science Foundation indicated that fewer than 15% of those currently employed in a STEM (Science, Technology, Engineering, and Mathematics) field in the United States of America are under the age of 30. You conduct a similar survey to determine whether a lower proportion of STEM field workers in your state are under the age of 30. After collecting a random sample of 500 STEM field workers in your state, you find that 67 are under the age of 30.

(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.

(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.

(c) Use the appropriate applet to generate a distribution of sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.

(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.

8) The United States Census Bureau determined that the mean annual household income for all households in the United States is $87,864 with a standard deviation of $21,542. To determine whether a different average applies to Dutchess County residents you collect a random sample of 14 Dutchess County residents and find an average income of $74,540 and a standard deviation of $25,059. The sample data were not strongly skewed.

(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.

(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.

(c) Use the appropriate applet to generate a distribution of simulated sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.

(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.
9) According to their maker, the different colors of Skittles candies (red, orange, yellow, green, and purple) are produced in equal quantities. However, they are packaged by weight and size, not color distribution. Your friend does not believe the colors are produced equally, rather, a larger proportion of yellow Skittles are produced and packaged. To test this claim you order 10 of the “Fun Size” bags of Skittles from different distribution centers. You then empty all the bags into one bowl and count the number of yellow skittles. You find that 16 of the 60 Skittles are yellow.

(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.

(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.

(c) Use the appropriate applet to generate a distribution of sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.

(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.

10) A group of National Fish and Wildlife researchers are collecting data on the bear population in Bearenstein National Park. It is estimated that there are 640 grizzly bears in the park. Recently, the researchers have noticed a parasite among several of the bears and need to determine whether more than 10% of the bear population has the parasite. The researchers found that of 29 bears surveyed, 16 were male and 13 were female, the average weight of the bears was 692.5 pounds, and 5 of the 29 bears have the parasite. The sample data were not strongly skewed.

(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.

(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.

(c) Use the appropriate applet to generate a distribution of sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.

(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.
11) The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all men ages 20-29 to be 5.8 feet with a standard deviation of 0.35 feet. You want to determine whether the average height of Dutchess County men ages 20-29 is greater than the national average. You survey a random sample of 95 Dutchess County residents between the ages of 20 and 29 that identify as men and calculate an average height of 5.91 feet and a standard deviation of 0.54 feet. The sample data were not strongly skewed.

(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.

(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.

(c) Use the appropriate applet to generate a distribution of sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.

(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.

12) The 2007-2008 United States National Health and Nutrition Examination Survey found the average height of all women ages 20-29 to be 5.375 feet with a standard deviation of 0.283 feet. You want to determine whether the average height of Dutchess County women ages 20-29 is different than the national average. You survey a random sample of 14 Dutchess County residents ages 20-29 that identify as women and calculate an average height of 5.2 feet and a standard deviation of 0.459 feet. The sample data were not strongly skewed.

(a) Write the null and alternative hypotheses in both words and symbols. Identify the sample size and label it using the appropriate symbol.

(b) Determine whether it would be appropriate to use a theoretical approach to conduct a statistical analysis.

(c) Use the appropriate applet to generate a distribution of sample statistics.
   a. If the criteria have been met use the Theory-based Inference applet.
   b. If the criteria have not been met to use the theoretical approach, explain why not. Using the One Proportion or the One Variable with Sampling applet, compare and contrast the simulated distribution with the normal distribution by using the “Normal Approximation” or “Overlay Normal Distribution” setting. Use complete sentences.
   c. Include a screenshot of the applet, distribution, and p-value found.

(d) Determine the appropriate formula (based on your variable type and approach) to calculate the standardized statistic. Write the formula symbolically, then show the calculation by substituting in the appropriate values, simplifying the expression using the appropriate order of operations, and rounding your answer to 3 decimals.

(e) Write a conclusion in context about the hypotheses using complete sentences and interpret the p-value in the context of the research question.
Comparing Two Means

The comparison of two population means is very common. Researchers may want to compare the average body temperature between male and female Chinstrap penguins, or average test scores between online and in-person classes. A key point to notice in both scenarios is that we first have a categorical variable that creates groups (sex, class type) and a quantitative variable that we measure (body temperature, test scores). In general, we will conduct an analysis by taking samples from each population and comparing them. More specifically, we will be asking:

"Are these population means the same?"

Moving forward, we will call the two general populations A and B. This leads to the following symbols:

<table>
<thead>
<tr>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter: $\mu_A$</td>
<td>Parameter: $\mu_B$</td>
</tr>
<tr>
<td>Statistics: $\bar{x}_A, s_A$</td>
<td>Statistics: $\bar{x}_B, s_B$</td>
</tr>
<tr>
<td>Sample Size: $n_A$</td>
<td>Sample Size: $n_B$</td>
</tr>
</tbody>
</table>

In general, we will use the following hypotheses:

$$H_0: \mu_A = \mu_B \quad \text{There is no difference between the two population means.}$$
$$H_A: \mu_A \neq \mu_B \quad \text{There is a difference between the two population means.}$$

Saying two numbers are equal is the same as saying there is no difference between the numbers. In other words, we could also write our hypotheses in the following form:

$$H_0: \mu_A - \mu_B = 0 \quad \text{There is no difference between the two population means.}$$
$$H_A: \mu_A - \mu_B \neq 0 \quad \text{There is a difference between the two population means.}$$

In some situations, it is more appropriate to do a ‘matched’ or ‘paired’ analysis. Consider a study in which you want to determine whether people like Pepsi or Coke more. In a matched pairs design, you get each person to rate and compare both sodas instead of splitting into two groups and rating only one soda. For a more thorough discussion of that type of study, see here$^1$. 

---

**Section Objectives**

- Use the Two means applet to create a distribution of simulated statistics for a difference in means.
- Use the Theory-based Inference applet to create a theoretical distribution of sample statistics for a difference in means.
- Verify that the validity conditions are met for using a theory-based approach.
- Conduct and interpret hypothesis tests for two population means.
Consider the following research question: Is the average amount of time (in hours) boys and girls ages seven to 11 spend playing sports each day the same?

(a) Describe the two populations of interest.
(b) Write out the hypotheses in both words and in symbols.

**Solution 3.4.1**

(a) The two populations are all boys ages seven to 11 and all girls ages seven to 11.

(b) 

\[ H_0: \mu_B = \mu_G \]  
There is no difference between the mean amount of time (in hours) all boys and all girls ages seven to 11 play sports each day.

\[ H_A: \mu_B \neq \mu_G \]  
There is a difference between the mean amount of time (in hours) all boys and all girls ages seven to 11 play sports each day.

We will rewrite this in a form more useful for simulation:

\[ H_0: \mu_B - \mu_G = 0 \]  
There is no difference between the mean amount of time (in hours) all boys and all girls ages seven to 11 play sports each day.

\[ H_A: \mu_B - \mu_G \neq 0 \]  
There is a difference between the mean amount of time (in hours) all boys and all girls ages seven to 11 play sports each day.

It is at this point in the previous chapters that we have created a distribution of simulated sample statistics based on the assumption that the null hypothesis is true. We will do the same thing here. For Example 3.4.1 we are assuming \( \mu_G = \mu_B \): the mean amount of time (in hours) that all boys play sports is the same as the mean amount of time (in hours) that all girls play sports.

What does \( \mu_G = \mu_B \) actually say? It says that the amount of time a child plays sports each day is not associated with their sex. In essence, we assume there is no connection between the two variables (no connection between sex and sports play time). To think about how to simulate this, consider the following set of flashcards.

Here are a few examples of what our data might look like:

- Boy 1: 1 hour
- Boy 2: 1.5 hours
- Boy 3: 3 hours
- Girl 1: 0.5 hour
- Girl 2: 2 hours
- Girl 3: 3.5 hours
Imagine I am carrying flashcards where each child is paired with a card that has their play time on it. I trip and fall over my dog Daisy’s toy (was this on purpose?).

All of my cards have fallen on the ground! I don’t remember which time went with which child so I randomly put them back together.

Now each child is randomly assigned a time from the set of all times in the sample. Boys may now be assigned to times that were originally girls’ time, and vice versa. If there was a difference in average daily play time between the groups, we have removed it by shuffling the times and reassigning them to sex!

We next compute the average daily play time in hours for each (sex) group and compare. This is our first simulation. We do this 100 or 1000 times to create a distribution and then find a p-value and/or standardized statistic.

Photo Credits:
https://unsplash.com/s/photos/open-door
To use the computer to do the simulation, we head to the Two means applet.

We copy the data and paste it into the box under ‘Enter data’ and then click the Use Data button. Next, check the “Show Shuffle Options” box. Finally, click the “Shuffle Responses” button.
The first shuffle leads to a difference in sample means of $-0.012$. That is, after the shuffle the average daily play time for boys minus the average daily play time for girls is $-0.012$ hours. In symbols, this is $\bar{x}_B - \bar{x}_G = -0.012$. The applet places a square at $-0.012$ on the distribution.

What is the meaning of a difference in sample means of $-0.012$ in this distribution?

**Interpretation:**
Under the assumption that there is really no difference between the average daily play time of all girls and all boys, this shuffle has the girl group with a sample mean that is $0.012$ hours higher than the mean of the boy group.

We click the ‘Shuffle Responses’ button again and get a difference in sample means of $0.03$ hours.

What is the meaning of a difference in sample means of $0.03$?

**Interpretation:**
Assuming there is no difference between the average daily play time of all girls and all boys, after this shuffle the mean of the boy group is $0.03$ hours higher than the mean of the boy group.

After 20 shuffles we see the distribution of sample statistics above. Even with only 20 squares we see a grouping of differences near $0$. We expect this since we are assuming that boys and girls have the same mean play time (i.e. the difference between the means is $0$). We then notice some variability (again we expect this!), in that some sample statistics are higher than $0$ and some are lower than $0$. None of the simulated sample statistics gets past $1.2$ hours on either side. Let’s jump up to 100 shuffles and see what happens.

We see from the distribution on the left that $0$ hours is the center (mean) and most of the sample statistics (differences in means) are located around that value. As in the previous sections, we can estimate a range of usual/expected sample statistics and unusual/unexpected sample statistics.

Based on this distribution of simulated sample statistics, we can say the following: assuming there is no difference between average daily play time of all boys and all girls, we expect the difference in sample means to be between about $-0.8$ and $0.8$ hours.
To find the observed statistic in this distribution and determine a p-value, we need to use the following information:

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Sample Mean</th>
<th>Sample Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girls</td>
<td>9</td>
<td>2</td>
<td>0.875</td>
</tr>
<tr>
<td>Boys</td>
<td>16</td>
<td>3.2</td>
<td>0.98</td>
</tr>
</tbody>
</table>

The statistic of interest is $\bar{x}_B - \bar{x}_G = 3.2 - 2 = 1.2$ hours. Since the alternative hypothesis uses $\neq$, we are running a two-sided test and must select “Beyond” from the “Count Samples” dropdown box. Entering the observed statistic in the Count box will provide us with a p-value:

```
Count Samples: Beyond, Count = 2/100 (0.0200)
```

In this case, p-value is 0.02 (different people running the simulation on different devices may get slightly different p-values). That is, assuming no difference in average daily play time of all boys and all girls, the probability of getting an observed statistic at least as extreme as 1.2 hours is 2%. Since the p-value is less than 5% the null hypothesis is not plausible. The alternative hypothesis is plausible.

We conclude that we have strong evidence that there is a difference in the mean daily play time between all girls and all boys ages seven to 11.
Use given data (or collect the same type of data from students in the class) to create flashcards to represent students in a class. Mark each pair with a number (or student initials) so they can be connected again in case of an accident. Each student should have two flashcards associated with them:

<table>
<thead>
<tr>
<th>One of</th>
<th>One of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Bird, Night Owl</td>
<td>Typical hours of sleep per night</td>
</tr>
</tbody>
</table>

Data:

(a) Compute the average typical nightly sleep for each group. Find the difference (Early Bird – Night Owl) between these averages and record it.

(b) Have the instructor very carefully collect these pairs. Make sure not to mess them up, we need them paired!

(c) (Instructor) Trip and fall. Try your best to really throw the flashcards all over the place.

(d) Have each student pick up two flashcards:

<table>
<thead>
<tr>
<th>One of</th>
<th>One of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Bird, Night Owl</td>
<td>Typical hours of sleep per night</td>
</tr>
</tbody>
</table>

(e) As a class, combine your results. Recompute the average nightly sleep for each group, find the difference (Early Bird – Night Owl) between these averages and record it.

(f) Discuss what has changed in the difference of means.

(g) (Optional) Carefully recombine the original pairs. Fall again. Repeat (d) – (f).

(h) Use the Two means applet to simulate 1000 shuffles of the data.

(i) Find the p-value for the observed statistic.

(j) Write a conclusion in context.
Comparing Two Means – Theory Based Approach

Note that in this section we are NOT assuming the two populations have equal standard deviations, but we are assuming the populations are independent. For a more thorough discussion of these topics see here.

Calculations by Hand
As stated previously, the comparison of two population means is very common. We are interested in understanding whether the mean of one population is different than the mean of a second population. In other words, we want to understand whether \( \mu_A = \mu_B \) (or whether \( \mu_A - \mu_B = 0 \)). We collect a sample from each population, calculate the mean from each sample, and then calculate the observed statistic, \( \bar{x}_A - \bar{x}_B \).

We then create a theoretical distribution of sample statistics (differences in sample means) that is centered at the null hypothesized value, \( \mu_A - \mu_B = 0 \). This distribution uses a mathematical model to predict the kinds of sample mean differences that would occur if the population means were the same. To calculate the standard error of this distribution of sample statistics, we also need to use the standard deviations from the samples, \( s_A \) and \( s_B \). The formula for the standard error and the formula for the standardized statistic (the t-score) are given below.

\[
\text{standard error} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{t - score} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

Since these formulas are somewhat more complicated than what we have seen previously, we will let a computer do the calculations for us! Note that, as with the t-distribution for a single mean, there are degrees of freedom associated with the t-distribution for a difference in two sample means. For more information, take a look at the Aspin-Welch t-test, here.

Calculations with Technology
In order to find the theory-based standard error, t-score, and p-value, we can use technology. The Theory-based Inference applet, using the “Two means” scenario, will allow us to find each of these values.

Caution – Validity Conditions
When the sample sizes of both Sample A and Sample B are five or larger and the sample group distributions are not strongly skewed, then the theory-based approach can be used.

If these conditions are not met, we should not use the theory-based approach and should instead use simulation.

2 https://openstax.org/books/introductory-statistics/pages/10-introduction
Example 3.4.2

The average amount of time all boys compared to all girls ages seven to 11 spend playing sports each day is believed to be the same. To explore this a study is done and data are collected, resulting in the summary statistics in the table below. The data are fairly symmetric in both groups.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Sample Mean</th>
<th>Sample Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girls</td>
<td>9</td>
<td>2</td>
<td>0.875</td>
</tr>
<tr>
<td>Boys</td>
<td>16</td>
<td>3.2</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Is there a difference in the mean amount of time all boys and girls ages seven to 11 play sports each day? Test this at the 5% level of significance.

Solution 3.4.2

We are interested in determining whether there is a difference in population means. Our null hypothesis says the average amount of time all boys compared to all girls ages seven to 11 spend playing sports each day is equal. Another way to say this is that there is no (zero) difference between the two population means.

\[ H_0: \mu_B - \mu_G = 0 \]  
There is no difference between the mean amount of time all boys and all girls ages seven to 11 play sports each day.

\[ H_A: \mu_G - \mu_G \neq 0 \]  
There is a difference between the mean amount of time all boys and all girls ages seven to 11 play sports each day.

Recall that the observed statistic is \( \bar{x}_B - \bar{x}_G = 1.2 \).

Since the size of each group is at least five and there is no strong skew in either group, we can use a theory-based approach to conduct the analysis. We will use the Theory-based Inference applet as shown below.
A study is conducted to determine whether Company A retains its workers longer than Company B. Company A samples 15 workers and their average time with the company is five years with a standard deviation of 1.2 years. Company B samples 20 workers and their average time with the company is 4.5 years with a standard deviation of 0.8 years. Neither of the samples was strongly skewed. Run the appropriate test using technology with a 5% level of significance. Don’t forget to state your final conclusion in context.
Section 3.4 | Comparing Two Means

Exercises

For questions 1 through 3, complete the following steps for the given scenario:

(a) Identify the observational units and state the sample sizes including the correct symbol.
(b) Identify the variables and classify them as categorical or quantitative.
(c) Identify the statistics of interest and include the appropriate symbol for each.
(d) Identify (describe in words) the parameters of interest and include the appropriate symbol for each.
(e) Write the null and alternative hypotheses in both words and symbols. Remember: When comparing two means, the null hypothesis is always the assumption that the two means are equal. That is, the difference between them is equal to zero.
(f) Use the Theory-based Inference applet to perform a statistical analysis. Include a screenshot of the applet that shows what you entered, the resulting distribution, standardized statistic, and p-value.
(g) Fill in the blanks for the statement interpreting the p-value.
(h) In complete sentences, write a conclusion regarding the research question.

1) You are moving from New York to Massachusetts and are trying to determine whether the average annual price of car insurance for adults ages 18-25 is the same in each state. You collect random samples of New York residents and Massachusetts residents between the ages of 18-25 and obtain the following statistics. The sample distributions were not strongly skewed.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Average Price of Car Insurance for 18- to 25-year-olds</th>
<th>Sample Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>33</td>
<td>$7,978</td>
<td>$986</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>25</td>
<td>$8,421</td>
<td>$702</td>
</tr>
</tbody>
</table>

Statement for part (g): If the average price of car insurance is the same in New York and Massachusetts for all 18 to 25-year-olds, the probability of obtaining a difference in sample statistics at least as extreme as ______ is ______%.

2) You tried two different planting methods for your vegetable garden this year. Half of your potatoes were planted using unaltered topsoil, while the other half were planted in soil which was supplemented with compost. You harvest the potatoes throughout the season, keeping them separate by planting method, then select a random sample from each group and measure their weights. The resulting summary statistics are shown below. The sample distributions were not strongly skewed. You wish to determine whether the compost supplement results in heavier (larger) potatoes.

<table>
<thead>
<tr>
<th>Soil Makeup</th>
<th>Sample Size</th>
<th>Average Weight</th>
<th>Sample Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular Soil + Compost Supplement</td>
<td>30</td>
<td>206 grams</td>
<td>93 grams</td>
</tr>
<tr>
<td>Regular Soil</td>
<td>30</td>
<td>162 grams</td>
<td>41 grams</td>
</tr>
</tbody>
</table>

Statement for part (g): If average weight of potatoes grown in a compost supplement is equal to the average weight of potatoes grown in unaltered topsoil the probability of obtaining a difference in average weights of ______ or more is ______%.
3) A group of National Fish and Wildlife researchers are collecting data on the grizzly bear population in Bearenstein National Park. It is estimated that there are 640 grizzly bears in the park. In a previous study the researchers found that there was evidence to suggest that a parasite was present in a significant portion of the population of grizzly bears in the park.

**Part 1:** The researchers are now concerned that the bears that are hosts to the parasite are losing weight at an unhealthy rate. The researchers obtain a sample of 93 bears at Bearenstein National Park and calculate the statistics below. The data from both groups were roughly symmetric.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Average Weight</th>
<th>Sample Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parasite</td>
<td>17</td>
<td>440 lbs.</td>
<td>89 lbs.</td>
</tr>
<tr>
<td>No Parasite</td>
<td>76</td>
<td>480 lbs.</td>
<td>120 lbs.</td>
</tr>
</tbody>
</table>

**Statement for part (g):** If the average weight of bears infected with the parasite is equal to the average weight of bears not infected with the parasite the probability of obtaining a difference in sample average weights at least as extreme as ______ is _____%.

**Part 2:** It is determined that the sex of the bear may be a confounding variable in the study. Define the term “confounding variable” and explain what it would mean for the sex of the bear to be a confounding variable in this scenario.

After determining that the sex of the bear needs to be accounted for in the study, the data was reviewed and organized into the following color-coded table:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Average Weight</th>
<th>Sample Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parasite</td>
<td>11</td>
<td>515 lbs.</td>
<td>130 lbs.</td>
</tr>
<tr>
<td>No Parasite</td>
<td>40</td>
<td>640 lbs.</td>
<td>88 lbs.</td>
</tr>
<tr>
<td>Female</td>
<td>6</td>
<td>310 lbs.</td>
<td>92 lbs.</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>361 lbs.</td>
<td>74 lbs.</td>
</tr>
</tbody>
</table>

**Research Question 1:** Is there a difference in the average weight of all male bears in the park infected with the parasite as compared to all male bears in the park that are not infected with the parasite?

**Complete steps (d) through (h) for Part 2, Research Question 1, including the following sentence for part (g):** Assuming that the average weight of all male bears infected with the parasite is equal to the average weight of all male bears not infected with the parasite the probability of obtaining a difference in sample average weights at least as extreme as ______ is _____%.

**Research Question 2:** Is there a difference in the average weight of all female bears in the park infected with the parasite as compared to all female bears in the park that are not infected with the parasite?

**Complete steps (d) through (h) for Part 2, Research Question 2, including the following sentence for part (g):** If the average weight of all female bears infected with the parasite is equal to the average weight of all female bears not infected with the parasite the probability of obtaining a difference in sample average weights at least as extreme as ______ is _____%.
Comparing Two Proportions

Running a hypothesis test to compare two population proportions is similar to running a hypothesis test to compare two means. Generally, the null hypothesis states that the two population proportions are the same, or that the difference between them is 0. That is,

\[ H_0: \pi_A = \pi_B \quad \text{or} \quad H_0: \pi_A - \pi_B = 0 \]

Simulation

In a typical scenario, we will have the following hypotheses:

\[ H_0: \pi_A = \pi_B \quad \text{There is no difference between the two population proportions.} \]
\[ H_A: \pi_A \neq \pi_B \quad \text{There is a difference between the two population proportions.} \]

Saying two numbers are equal is the same as saying there is no difference between the numbers, which means the hypotheses can also be written in a slightly different form:

\[ H_0: \pi_A - \pi_B = 0 \quad \text{There is no difference between the two population proportions.} \]
\[ H_A: \pi_A - \pi_B \neq 0 \quad \text{There is a difference between the two population proportions.} \]

The simulation approach will mirror the work done in Section 3.4 with comparing two means. We will again have two groups which are created based on a categorical variable. In this section, though, the second variable is also categorical. This leads us to a difference in proportions as the observed statistic, \( \hat{p}_A - \hat{p}_B \).

Again, we will calculate simulated sample statistics under the assumption that the null hypothesis is true. These simulated statistics will form a null distribution. We can then determine whether the observed sample statistic is a usual/expected or unusual/unexpected value under the assumption of the null hypothesis (no difference between the population proportions).
A pharmaceutical company has developed a new medication to treat hives. They would like to determine whether the new medication does a better job of treating hives than the current leading medication. The company randomly selects 400 adults suffering from hives and randomly assigns half of them to take medication A (the new drug) and half to take medication B (the current leading drug). Of the 200 subjects taking medication A, 20 still had hives. Of the 200 subjects taking medication B, 12 still had hives.

(a) Describe the two populations and parameters of interest.
(b) Write the null and alternative hypotheses in both words and symbols.

**Solution 3.5.1**

(a) Population A is the set of all adults with hives who take medication A. Population B is the set of all adults with hives who take medication B. \( \pi_A \) is the proportion of all adults who still have hives after taking medication A. \( \pi_B \) is the proportion of all adults who still have hives after taking medication B.

(b) The null hypothesis says the proportions described above are equal. Another way to say this is that there is no (zero) difference between the two proportions.

\[
H_0: \pi_A = \pi_B \quad \text{The proportion of all adults who still have hives after taking medication A is the same as the proportion of all adults who still have hives after taking medication B.}
\]

\[
H_A: \pi_A \neq \pi_B \quad \text{The proportion of all adults who still have hives after taking medication A is not the same as the proportion of all adults who still have hives after taking medication B.}
\]

Or rewritten:

\[
H_0: \pi_A - \pi_B = 0 \quad \text{There is no difference between the proportion of all adults who still have hives after taking medication A and the proportion of all adults who still have hives after taking medication B.}
\]

\[
H_A: \pi_A - \pi_B \neq 0 \quad \text{There is a difference between the proportion of all adults who still have hives after taking medication A and the proportion of all adults who still have hives after taking medication B.}
\]
The process for creating a distribution of simulated sample statistics is the same as for two means. In order to simulate no association between the medication you take (A vs B) and whether your hives clear up, we shuffle! For this example, we can again create flashcards to represent the data.

Below are a few examples of what the data look like:

We would do this by having 200 cards say Medication A and 200 cards say Medication B. We then take our 368 No Hives and 32 Hives cards and randomly pair them with the medication cards. We might expect that half of the Hives cards (16) would go to the Medication A group while the other half (16) would go to the Medication B group. This would give us a sample statistic of \( \hat{p}_A - \hat{p}_B = \frac{16}{200} - \frac{16}{200} = 0 \).

On the next shuffle, we might have 14 Hives cards go to the Medication A group and 18 Hives cards go to the Medication B group. In this case, the simulated statistic is \( \hat{p}_A - \hat{p}_B = \frac{14}{200} - \frac{18}{200} = -0.02 \).

Of course, it would take a lot of time and work to repeat this shuffling, counting, and calculation process 1000 times. Instead we will have the computer help us.

We can use the Two-way Tables applet:
We can then paste the data and click the “Use Data” button (left screenshot below) or check the “Enter table” box, enter summary statistics for each group, and click the “Use Table” button (right screenshot below). Note that we can also enter names with no spaces for each explanatory variable group (MedA, MedB) and for the success and failure outcomes (Hives, NoHives). In this case, since researchers want to keep track of whether subjects still have hives after taking the medication, having hives is considered the “Success” outcome!

Next, we check the “Show Shuffle Options” box and input the “Number of Shuffles”, in this case 1000:

Look at the output under “Most Recent Shuffle” on the left side of the screenshot above. The applet has color coded cards according to the outcomes of the response variable. There are 32 blue cards denoting “Success” (subject still has hives) and 368 green cards denoting “Failure” (subject does not have hives). For each repetition, all 400 cards are shuffled and randomly dealt into two piles of 200, since there were 200 subjects in each explanatory variable (MedA and MedB) group. In the most recent shuffle, 16 subjects with hives wound up in each medication group, and the resulting simulated statistic was $\hat{p}_A - \hat{p}_B = \frac{16}{200} - \frac{16}{200} = 0$.

The distribution of simulated sample statistics is shown at right in the screenshot above. Clicking on any of the squares in the distribution will show the group counts for that particular repetition under “Most Recent Shuffle”. Notice that the distribution is bell-shaped and is centered at 0 (mean = 0.002), which matches the null hypothesis!
Example 3.5.2

A pharmaceutical company has developed a new medication to treat hives. They would like to determine whether the new medication does a better job of treating hives than the current leading medication. The company randomly selects 400 adults suffering from hives and randomly assigns half of them to take medication A (the new drug) and half to take medication B (the current leading drug). Of the 200 subjects taking medication A, 20 still had hives. Of the 200 subjects taking medication B, 12 still had hives. Determine how likely it would be to see a difference at least as extreme as 0.04 assuming there is no difference between the medications.

Solution 3.5.2

From our work in Example 3.5.1 we have the following hypotheses:

\[ H_0: \pi_A - \pi_B = 0 \quad \text{There is no difference between the two population proportions.} \]
\[ H_A: \pi_A - \pi_B \neq 0 \quad \text{There is a difference between the two population proportions.} \]

The observed statistic is: \( p_A - p_B = \frac{20}{200} - \frac{12}{200} = 0.10 - 0.06 = 0.04 \), which is shown in the applet to the left of center near the bottom.

\text{Observed Diff(Hives)= 0.040}

Continuing with the Two-way Tables applet, we use the Count box to find the requested probability. Since the alternative hypothesis involves the \( \neq \) symbol, we choose “Beyond” from the dropdown box next to “Count Samples”. We then enter the observed statistic in the Count box and click the “Count” button.

The p-value is 0.173 which means it would not be unusual to see a difference in the proportion of subjects who still have hives of at least 0.04 if there was really no difference between the medications. This leads us to conclude that the null hypothesis is plausible. That is, it is plausible that there is no difference between the proportion of all adults who still have hives after taking medication A and the proportion of all adults who still have hives after taking medication B. There is no statistically significant evidence that the new medication is more effective than the old medication.
A pharmaceutical company has developed a new medication to treat hives. They would like to determine whether the new medication does a better job of treating hives than the current leading medication. The company randomly selects 400 adults suffering from hives and randomly assigns half of them to take medication A (the new drug) and half to take medication B (the current leading drug). Of the 200 subjects taking medication A, 20 still had hives. Of the 200 subjects taking medication B, 12 still had hives. Use a theory-based approach to determine how likely it would be to see a difference at least as extreme as 0.04 assuming there is no difference between the medications. Use a significance level of 0.01 to make a conclusion in context.

**Solution 3.5.3**

From our work in Example 3.5.1 we have the following hypotheses:

\[ H_0: \pi_A - \pi_B = 0 \quad \text{There is no difference between the two population proportions.} \]

\[ H_A: \pi_A - \pi_B \neq 0 \quad \text{There is a difference between the two population proportions.} \]

The observed statistic is:

\[ p_A - p_B = \frac{20}{200} - \frac{12}{200} = 0.10 - 0.06 = 0.04. \]

Because there are at least 5 successes and 5 failures in each group and the population is at least 10 times the size of the sample (at least 10*400 = 4000), we can use the **Theory-based Inference** applet. We choose the “Two proportions” scenario from the dropdown box at the top left and fill in the summary data for each group.

We then check the “Test of significance” box at the top right, enter the value 0 from the null hypothesis, change the symbol for the alternative hypothesis to \( \neq \), and click the “Calculate” button.

As we can see from the screenshot at left, the theory-based p-value is 0.1404. Since this is greater than the significance level (0.01) there is little to no evidence against the null hypothesis. The null hypothesis is plausible.

We conclude that it is plausible that there is no difference in the proportions of all adults who still have hives when taking Medication A versus Medication B.

This matches the conclusion using simulation!
Comparing Two Proportions – Theory Based Approach

When conducting a theory-based hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The number of successes is at least five, and the number of failures is at least five, for each of the sample groups.
2. The population should be at least 10 times the size of the sample.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a true difference in the population proportions or it may be due to chance. A hypothesis test can help determine whether a difference in sample proportions is plausible under the “no difference in the population proportions” assumption. The difference in two sample proportions follows an approximately normal distribution. As we have seen, the null hypothesis states that the two proportions are the same. That is, \( H_0: \pi_A = \pi_B \) or \( H_0: \pi_A - \pi_B = 0 \). To conduct the test, we use a pooled proportion of successes, \( \hat{p}_c \). Given the number of successes for samples A and B, \( x_A \) and \( x_B \) respectively, we calculate the pooled proportion of successes as follows:

\[
\hat{p}_c = \frac{x_A + x_B}{n_A + n_B}
\]

For the previous scenario, we calculate \( \hat{p}_c = \frac{20 + 12}{200 + 200} = \frac{32}{400} = 0.08 \).

The formula for the standard error is \( SE = \sqrt{\hat{p}_c(1 - \hat{p}_c) \left( \frac{1}{n_A} + \frac{1}{n_B} \right)} \). For the previous scenario, we calculate \( SE = \sqrt{0.08(1 - 0.08) \left( \frac{1}{200} + \frac{1}{200} \right)} = 0.027 \). Note that this matches the SD in the theory-based distribution shown in Example 3.5.3.

The formula for the standardized statistic (z-score) is \( z = \frac{(\hat{p}_a - \hat{p}_b) - 0}{\hat{p}_c(1 - \hat{p}_c) \left( \frac{1}{n_A} + \frac{1}{n_B} \right)} \). The formula looks complicated, but it is still the same formula that we saw in Section 2.3 for the standardized statistic. That is,

\[
z = \frac{\text{observed statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}
\]

For the previous scenario, we calculate \( z = \frac{0.04 - 0}{0.027} = 1.48 \).

This is slightly different than the standard score given by the applet (1.47). The difference is due to rounding error because we used a rounded value (0.027) for the standard error.

Caution – Validity Conditions

Using the theory-based approach is comparable to the simulation-based approach as long as:

- Each sample group has at least 5 successes and 5 failures.
- The population is at least 10 times the size of the sample.

If these conditions are not met, we should not use the theory-based approach and should instead use simulation.
Section 3.5 | Comparing Two Proportions

Exercises

For questions 1 through 3, complete the following steps for the given scenario:

(a) Identify the observational units and state the sample sizes including the correct symbol.
(b) Identify the variables and classify them as categorical or quantitative.
(c) Identify the statistics of interest and include the appropriate symbol for each.
(d) Identify (describe in words) the parameters of interest and include the appropriate symbol for each.
(e) Write the null and alternative hypotheses in both words and symbols. Remember: When comparing two proportions, the null hypothesis is always the assumption that the two proportions are equal. That is, the difference between them is equal to zero.
(f) Use the appropriate applet (either the Two-way Tables applet or the Theory-based Inference applet) to perform a statistical analysis. Determine the values and symbols that should be entered in the applet to create a distribution of simulated sample statistics. Use the applet to calculate the p-value and the standardized statistic. Include a screenshot of the applet that shows what you entered, the resulting distribution, standardized statistic, and p-value.
(g) Fill in the blanks for the statement interpreting the p-value.
(h) In complete sentences, write a conclusion regarding the research question.

1) An advertisement for Supplement A (a popular vitamin supplement) states that it helps fight off colds faster and can even prevent colds if you take the supplement “at the FIRST sign of a cold symptom or before entering crowded environments”. However, you also see a warning message that “These statements have not been evaluated by the Food and Drug Administration. This product is not intended to diagnose, treat, cure, or prevent any disease”. You decide to conduct your own study to determine whether Supplement A is worth using during the next cold season.

You select a random sample of 75 people ages 18 to 65 and provide 36 with several doses of Supplement A to use as stated in the directions, and 39 with a placebo pill and the same instructions. After 8 weeks you conduct a follow-up interview and record whether they had a cold during the 8 weeks. A summary of the results is shown in the table below.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Did Not Experience the Common Cold Within 8 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplement A</td>
<td>36</td>
</tr>
<tr>
<td>Placebo Pill</td>
<td>39</td>
</tr>
</tbody>
</table>

Statement for part (g): Assuming Supplement A is not effective at preventing colds, the chance of seeing a difference in the proportion of participants in each group who did not experience a cold at least as extreme as _____ is _____%.
2) Many popular shoe companies once produced “toning sneakers”, which they claimed would strengthen hamstrings and calves and tone the buttocks more than regular sneakers. You construct a study to test the claim “Toning sneakers are just as effective as regular sneakers at strengthening and toning leg muscles”.

**Part 1:** A random sample of 80 people ages 18 to 35 are randomly assigned to two groups. In Group A, you provide everyone with a pair of toning sneakers. In Group B, you provide everyone with a pair of regular sneakers of the same brands. The participants agree to wear the sneakers for the next month and return for a follow-up interview. During the initial interview and the follow-up interview, you conduct a series of leg strength exercises and measure the maximum number of repetitions as well as note whether there is a visible change in the appearance of “muscle tone”. A summary of the results is shown in the table below.

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample Size</th>
<th>Statistically Significant Change in Strength and Muscle Tone</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td>B</td>
<td>38</td>
<td>7</td>
</tr>
</tbody>
</table>

**Statement for part (g):** Assuming toning sneakers are just as effective as regular sneakers, the chance of seeing a difference in the proportions of participants in each group that experienced a change in strength and muscle tone that is at least as extreme as ______ is ______%.

**Part 2:** One of the follow-up interview questions was “Have you made any major diet and/or exercise changes in the past 30 days?” Those who responded “Yes” were removed from the study to attempt to only look at the effect of the sneaker itself (and not any additional changes that occurred in their lives due to being in a study regarding health and fitness).

Below are the summary results from the study with the participants who changed their diet/exercise removed.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Statistically Significant Change in Strength and Muscle Tone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>Group B</td>
<td>35</td>
<td>6</td>
</tr>
</tbody>
</table>

**Statement for part (g):** Assuming toning sneakers are just as effective as regular sneakers, the chance of seeing a difference in the proportions of participants in each group that experienced a change in strength and muscle tone that is at least as extreme as ______ is ______%.
3) A group of National Fish and Wildlife researchers are collecting data on the grizzly bear population in Bearenstein National Park. It is estimated that there are 640 grizzly bears in the park. In a previous study the researchers found that there was evidence to suggest that a parasite was present in a significant portion of the population of grizzly bears in the park, and in Section 3.4 we determined that the parasite is causing significant weight change in the male grizzly population.

Additional research is now required on the effects of the parasite on the bear population. Specifically, researchers want to determine how the parasite is spread and believe that this question can be partially answered by looking at the infection rates across sexes of the bear population.

The researchers collected a new random sample of 105 bears and noted the sex of each bear, as well as whether the parasite was present. A summary of the results is shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parasite</td>
<td>22</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>No Parasite</td>
<td>83</td>
<td>32</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>47</td>
<td>58</td>
</tr>
</tbody>
</table>

**Statement for part (g):** Assuming the parasite infects male and female bears at equal rates, the probability of obtaining a difference in infection rates at least as extreme as _____ is _____%.

4) You conduct a simple random sample to collect information on health insurance coverage from residents in your neighborhood, which has a population of approximately 2000 people. You recently read that the higher the household income, the more likely a person in that house is to be covered by employer-sponsored health insurance. A summary of the results is shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Employer-Sponsored Health Insurance</th>
<th>Uninsured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household Income below $75,000</td>
<td>86</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>Household Income greater than or equal to $75,000</td>
<td>64</td>
<td>50</td>
<td>3</td>
</tr>
</tbody>
</table>

**Statement for part (g):** Assuming the same proportion of people receive employer-sponsored health insurance regardless of their household income level, the probability of obtaining a difference in the proportion of households receiving employer-sponsored health insurance at least as extreme as _____ is _____%.
5) A migraine is a particularly painful type of headache that people sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a study where 89 New York residents who were experiencing migraine headaches were either given acupuncture treatment designed to treat migraines, or they were given a placebo treatment (using acupuncture, but in areas not intended to treat migraines). In a follow-up appointment the next day, participants were asked if they were now “pain-free”. A summary of the results is shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Pain-Free After 24 Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migraine Treatment</td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td>Placebo Treatment</td>
<td>46</td>
<td>7</td>
</tr>
</tbody>
</table>

**Statement for part (g):** Assuming acupuncture is not effective for treating migraines, the probability of obtaining a difference in the proportion of pain-free participants at least as extreme as _____ is _____%.
Chapter 4 | Confidence Intervals

Contents of Chapter 4

Section 4.1  Introduction to Confidence Intervals

Section 4.2  Theory-Based Approach to Confidence Intervals

Section 4.3  Impacts on the Width of a Confidence Interval
Recall that the **significance level**, \( \alpha \) (alpha), determines how small a p-value needs to be to provide convincing evidence to **reject** the null hypothesis.

- **a p-value larger than the significance level means we DO NOT have enough evidence to reject the null hypothesis.** We conclude that the null hypothesis IS plausible.
  
  **Example:** If we decide to use a significance level of \( \alpha = .05 \), then we would not reject the null hypothesis if the p-value is larger than 0.05 = 5%. For example, if the p-value is 0.068 (6.8%) or 0.052 (5.2%), then we would NOT reject the null hypothesis. We would conclude that the null hypothesis is plausible.

- **a p-value smaller than the significance level means we DO have enough evidence to reject the null hypothesis.** We conclude that the null hypothesis IS NOT plausible.
  
  **Example:** If we decide to use a significance level of \( \alpha = .05 \), then we would reject the null hypothesis if the p-value is smaller than 0.05 = 5%. For example, if the p-value is 0.024 (2.4%) or 0.049 (4.9%), then we WOULD reject the null hypothesis and conclude that the null hypothesis is not plausible.
We want to know whether a new drug reduces the proportion of people who suffer from a certain side effect. The current drug creates this side effect in 8% of users. We select a random sample of 240 users of the new drug and find that 5.2% of them suffer from the side effect. We decide to use a significance level of 0.01 to test whether the side effect rate for all users of the new drug is less than 8%. The distribution of simulated sample statistics and p-value are shown below.

(a) Do we reject or fail to reject the null hypothesis? Explain.
(b) Write a conclusion in context.

Solution 4.1.1

(a) We fail to reject the null hypothesis. The p-value is 0.045, which is larger than the significance level of 0.01.

(b) It is plausible that the proportion of all people who experience the side effect with the new drug is 8% (the same as the current drug). Therefore, we do not have evidence that the new drug reduces the proportion of people who suffer from the side effect.
Confidence Intervals

Up to this point, our hypothesis testing procedure has followed the pattern below:

<table>
<thead>
<tr>
<th>General Procedure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>State the hypotheses and identify the statistic(s) from the sample</td>
<td>( H_0 : \pi = 0.08, \ H_A : \pi \neq 0.08, \ \hat{p} = 0.05, \ \alpha = 0.01 ) where ( n = 240 )</td>
</tr>
<tr>
<td>Use ( H_0 ) to create a distribution of simulated sample statistics</td>
<td>Summarize Stats</td>
</tr>
<tr>
<td>Compare the observed statistic to this distribution by finding the p-value or standardized statistic.</td>
<td>Choose statistic: - Number of successes + Proportion of successes</td>
</tr>
<tr>
<td>Make conclusion: Reject the null or null is plausible.</td>
<td>Compare the observed statistic to this distribution by finding the p-value or standardized statistic.</td>
</tr>
<tr>
<td>The p-value (0.10) is larger than alpha (0.01). We conclude that the null hypothesis plausible. It is plausible that ( \pi = 0.08 ).</td>
<td></td>
</tr>
</tbody>
</table>

From this work we conclude that \( 0.08 = 8\% \) is a plausible value for the parameter. But what if we had hypothesized that \( \pi = 7\% \) or \( \pi = 9\% \)? Based on our sample statistic, could we also conclude that those are plausible values of the population parameter? The answer to this question is given by a confidence interval.

An Interval of Plausible Values

A confidence interval is an interval of plausible values for the population parameter. It can be written in two forms:

- (Observed Statistic) ± (Margin of Error)
- Interval notation: (lowest number in confidence interval, highest number in confidence interval)

A confidence interval will always have the observed statistic at its midpoint. This means that the observed statistic is always a plausible value of the population parameter since it will always be a value in the confidence interval. All of the values within the confidence interval are also plausible values for the population parameter. For example, if the confidence interval is \( 0.35 \pm 0.02 \), then this means that values ranging from \( 0.35 - 0.02 = 0.33 \) all the way up to \( 0.35 + 0.02 = 0.37 \) are plausible values of the population parameter. This confidence interval could also be written \( (0.33, 0.37) \) to indicate that the range of plausible values spans from 0.33 up to 0.37. This tells us that any number within that range is a plausible value of the population parameter. For example, the numbers 0.33, 0.34, 0.35720, 0.3601, and 0.37 are all plausible values of the population parameter.

One way to construct a confidence interval is by running multiple hypothesis tests to determine whether each value in the null hypothesis is plausible. Each plausible value will be included in the confidence interval, and each rejected value will not be included in the confidence interval.
The observed statistic will always be a plausible value for the parameter when conducting a hypothesis test. The observed statistic is always the midpoint of the confidence interval. We need to test values above and below the statistic in order to determine other plausible values for the parameter. This means we need to reject two values, one below the observed statistic and one above, before our full confidence interval is identified.

Continuing with Example 4.1.1, let’s try a few different values for the parameter and see what happens! Recall that the observed statistic, \( \hat{p} = 0.052 \), is the percentage of people in the sample who experience a certain side effect with the new drug. We are interested in estimating the percentage of all users of the new drug who experience the side effect. Let’s first set up a number line centered around the observed statistic to organize our possible plausible parameter values.

We start by testing the observed statistic, \( \hat{p} = 0.052 \). That is, we will assume \( \pi = 0.052 \) and run a two-sided hypothesis test to determine whether \( \pi = 0.052 \) is a plausible value of the population parameter. We are testing whether it is plausible to conclude that 5.2% of all users of the new drug experience the side effect.

Let’s mark this plausible value for the parameter on our number line with a purple dot:

The p-value is 1, which is bigger than 0.01 (\( \alpha \)). We conclude that the null hypothesis is plausible. This means we include 0.052 as a plausible value for \( \pi \).

Let’s mark this plausible value for the parameter on our number line with a purple dot:

Thus, based on our sample, 5.2% is a plausible value for the proportion of all users who experience the side effect while taking this new drug. How about 4% or 6%? Are these plausible values for \( \pi \)? Would we come to the same conclusion about values of \( \pi \) such as \( \pi = 2\% \) or \( \pi = 10\% \)? Is it plausible that 10% of all users experience the side effect? Is it plausible that 2% of all users experience the side effect? How high and how low do we need to go to capture the plausible values of the parameter?

The following screenshots show null distributions from hypothesis tests conducted for each of the values mentioned above. In each case, we are assuming a value for \( \pi \) and seeing whether the observed statistic (\( \hat{p} = 0.052 \)) is usual or unusual under that assumption. By comparing the p-value to \( \alpha = 0.01 \), we can then make a conclusion about the plausibility of each (null) hypothesized value of \( \pi \).
For each hypothesis test (including those with distributions pictured above), we compare the p-value to the significance level, \( \alpha = 0.01 \), in order to determine whether the hypothesized value of \( \pi \) will be contained in the confidence interval (because it is plausible) or not contained in the confidence interval (because it is not plausible).

- The values 0.04 and 0.06 have p-values larger than 0.01, and so they are plausible values of \( \pi \). Therefore, 0.04 and 0.06 are included in the confidence interval (marked in purple below).

- The values 0.02 and 0.10 have p-values less than 0.01, and so we reject them as plausible values of \( \pi \). Therefore, 0.02 and 0.10 are not included in the confidence interval (marked in red below).

Notice that every value between 0.04 and 0.06 has also been marked (in purple) as plausible. If 0.06 is not far enough away from 0.052 to be rejected, then any number in between them surely won’t be either!

So far, we have an interval of plausible values from 0.04 to 0.06. We can write this in interval notation as (0.04, 0.06). This means that, so far, we know that it is plausible that between 4% and 6% of all users of the new drug experience the side effect. We then continue this process of repeatedly testing values for \( \pi \), and adding them to our interval if they are plausible.

We construct another confidence interval for this scenario in the next example, this time using a different significance level.
Example 4.1.2

We want to estimate the value of a parameter, \( \pi \), the proportion of all users of a new drug that experience a certain side effect. The current drug creates this side effect in 8% of all users. We select a random sample of 240 users of the new drug and find that 5.2% of them suffer from the side effect. Using a significance level of 0.05, construct a confidence interval to estimate the value of the parameter.

Solution 4.1.2

We will start by testing our observed statistic, that is, \( H_0 : \pi = 0.052 \). From there we will use a two-sided test to test values both bigger and smaller than 0.052 until we reject a value on each side. We summarize this information in the table below.

Starting from \( H_0 : \pi = 0.052 \), we tested values above and below until we eventually rejected a value on both sides. To write this in interval notation, we use the last plausible value above and below 0.052 that we found, 0.082 and 0.032 respectively. Thus, the confidence interval is

\[ \pi : (0.032, 0.082) \]

Note that the interval must be written with the lower value on the left, and the higher value on the right.

We estimate that it is plausible that the proportion of all people who suffer from a certain side effect when taking this new drug is between 3.2% and 8.2%.
It is important to note that a confidence interval is made up of values for the parameter and not made up of p-values. The p-value is only a tool for deciding whether to reject a particular (null) hypothesized value. It is a common error for students to accidentally mix these values up when creating a confidence interval. In the above example, it would be incorrect to state the confidence interval, using the p-values from the table, as (0.071, 0.062) or as (0.062, 0.071).

**Confidence Level**

In Example 4.1.2, we decided to use a significance level of $\alpha = 0.05$. This was a choice and with a different choice we will get a slightly different confidence interval.

- For a significance level of $\alpha = 0.05$, the null hypothesis is plausible if the observed statistic falls within the middle 95% of the corresponding distribution of sample statistics.

- If we were to change the significance level to $\alpha = 0.01$, the null hypothesis would be plausible if the observed statistic falls within the middle 99% of the corresponding distribution of sample statistics.

The percentages, 95% and 99%, in the examples above are called confidence levels. When writing contextual conclusions, we need to include the confidence level in the statement. Below is a generic interpretation of a confidence interval, with different colors to indicate the pieces that need to be filled in for a particular scenario.

“We are (confidence level)% confident that the (population parameter) is between (confidence interval)”.

The statement below is an interpretation in context for the confidence interval constructed in Example 4.1.2.

“We are 95% confident that the proportion of all users of the new drug who suffer from a certain side effect is between 4.2% and 8.7%”.

If we write the confidence level and the significance level in decimal form, then each of the following statements is true:

- $(\text{Confidence Level}) = 1 - (\text{Significance Level})$
- $(\text{Significance Level}) = 1 - (\text{Confidence Level})$
- $(\text{Confidence Level}) + (\text{Significance Level}) = 1$

The most commonly used significance levels and their corresponding confidence levels are given in the table below.

<table>
<thead>
<tr>
<th>Significance Level, $\alpha$</th>
<th>Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>99% = 0.99</td>
</tr>
<tr>
<td>0.05</td>
<td>95% = 0.95</td>
</tr>
<tr>
<td>0.10</td>
<td>90% = 0.90</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$(1 - \alpha) \times 100% = 1 - \alpha$</td>
</tr>
</tbody>
</table>

In the next example we consider three confidence intervals with varying confidence levels.
Example 4.1.3

A pizza restaurant wants to estimate the average (mean) time it takes to deliver a pizza. To this end they take a random sample of 28 deliveries and find a sample mean of 36 minutes and a sample standard deviation of 7 minutes. The following are confidence intervals created with confidence levels of 90%, 95%, and 99%:

- 90%: (33.75, 38.25)
- 95%: (33.29, 38.71)
- 99%: (32.33, 39.67)

(a) Interpret each of the intervals in context.
(b) Would we believe the restaurant if it claimed to have an average delivery time of less than 40 minutes?
(c) What if their claim was less than 35 minutes?

Solution 4.1.3

(a) The given intervals consist of plausible values for the population mean, $\mu$, which is the average (mean) time it takes to deliver all pizzas from this restaurant. We use the generic interpretation template to write the contextual interpretations:

- 90%: We are 90% confident that the average delivery time for all pizzas is between 33.75 and 38.25 minutes.
- 95%: We are 95% confident that the average delivery time for all pizzas is between 33.29 and 38.71 minutes.
- 99%: We are 99% confident that the average delivery time for all pizzas is between 32.33 and 39.67 minutes.

(b) If the restaurant claims that their average (mean) delivery time is less than 40 minutes, then we would believe that claim since all of the confidence intervals have an upper limit that is less than 40. No matter which confidence interval we use, all of the plausible values are less than 40 minutes as the restaurant claimed. Therefore, the company’s claim that their average delivery time is under 40 minutes is believable.

(c) If the restaurant claims that their average delivery time is less than 35 minutes, then we would not believe this claim. All of the confidence intervals contain plausible values for $\mu$ that are larger than 35 minutes. Therefore, it is plausible that their average delivery time is longer than 35 minutes.

Margin of Error

In the previous examples we have written our confidence intervals in interval notation:

(lowest number in confidence interval, highest number in confidence interval)

A very useful alternative notation is to write the interval as:

$$(\text{Observed Statistic}) \pm (\text{Margin of Error})$$

You will see confidence intervals presented this way frequently in the media. In fact, it is not uncommon for the media to present the observed statistic as if it is the parameter, and to give the margin of error in the fine print. It is very important that we pay attention to both the statistic and the margin of error when digesting statistics ‘in the wild’.
We can go back and forth between the two confidence interval forms to suit our needs. In Example 4.1.3, we were given a 95% confidence interval for average pizza delivery time of (33.29, 38.71). To express this in the alternate notation we need to find the statistic and margin of error (MOE).

In order to find the observed statistic, remember that the observed statistic is always the midpoint of the confidence interval. To find the midpoint, we can find the mean of the endpoints of the confidence interval:

\[
\frac{33.29 + 38.71}{2} = 36 \text{ minutes} = \bar{x} = \text{midpoint of confidence interval}
\]

The number 36 is in the middle of the values 33.29 and 38.71. The number 36 is the original observed statistic, the average delivery time for the sample of 28 pizza deliveries. It is the value at the midpoint of the confidence interval.

To find the margin of error (MOE) we need to find the distance from the center to one of the ends of the interval:

\[
36 - 33.29 = 2.71
\]

or

\[
38.71 - 36 = 2.71
\]

The margin of error is 2.71 minutes.

We can now write the confidence interval as \( \mu: 36 \pm 2.71 \) or as \( \mu: (33.29, 38.71) \).

- We are 95% confident that the average delivery time for all pizza deliveries from this restaurant \( \mu \) is 36 minutes \( \pm 2.71 \) minutes.

- We are 95% confident that the average delivery time for all pizza deliveries from this restaurant \( \mu \) is between 33.29 minutes and 38.71 minutes.
Example 4.1.4

A Marist poll taken from October 29 to November 1, 2020, states that 51% of likely voters in Pennsylvania would support Joe Biden if the election were held that day. In fine print, the margin of error is given as 4.4%.

(a) Create a 95% confidence interval from the above information. Interpret the interval in context.

(b) Based on the interval, are we confident that Joe Biden would win a majority of votes (i.e., more than 50%) in Pennsylvania if the election were held that day?

Solution 4.1.4

(a) To create the confidence interval, we can start with the form (Observed Statistic) ± (Margin of Error). As long as we include the % symbol, the interval can be written as 51% ± 4.4%. Alternatively, we could convert the values to decimals and write the interval as 0.51 ± 0.044. To write the interval in interval notation, we can perform the following calculations:

\[
51\% \pm 4.4\%
\]

In decimal form, the confidence interval could also be written as \((0.466, 0.555)\).

To interpret the interval, we can use the generic interpretation template: “We are (confidence level)% confident that the (population parameter) is between (confidence interval)”. Filling in the pieces for this scenario, the interval can be interpreted as follows:

“We are 95% confident that the proportion of all likely voters in Pennsylvania who would vote for Joe Biden if the election were held that day is between 46.6% and 55.5%”.

Note that although a confidence level was not given in the polling report, the significance level for polls of this sort can be assumed to be 5% unless stated otherwise.

(b) Based on the confidence interval, we are not confident that Joe Biden would win a majority of votes in Pennsylvania if the election were held that day. Since the confidence interval includes plausible values that are less than 50%, we cannot say with confidence that Joe Biden would win more than 50% of votes in Pennsylvania if the election were held that day.
Section 4.1 | Confidence Intervals

Exercises

1) To estimate the average height of all Dutchess County residents ages 20 to 29 who identify as women, a random sample of 110 Dutchess County residents in that age group is selected. Of the 110 sampled, 36 identified as women. A 95% confidence interval is created for the average height of all Dutchess County residents ages 20 to 29 who identify as women: (5.3892, 5.5808) (or between 5’5” and 5’6”).
   a) What are the sample mean height and the margin of error for this confidence interval?
   b) The United States National Health and Nutrition Examination Survey claims that the average height of females ages 20 to 29 is 5.375 feet (5 feet 4.5 inches). Based on the confidence interval, is this a plausible value for the average height of all Dutchess County residents ages 20 to 29 who identify as women? Explain.
   c) Complete the following statement:
      We are ____% confident that the _______________ is between __________ and __________.

• The results of a random telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?”. Twenty percent answered “crime”. A 99% confidence interval is created to estimate the proportion of all adult Americans who feel that crime is the main problem facing the country. The confidence interval is 21% ± 3.3%
   a) Rewrite the confidence interval in interval notation.
   b) Your friend claims that one-quarter of all adult Americans believe that crime is the main problem facing the country. Is your friend’s claim plausible? Explain.
   c) Complete the following statement:
      We are ____% confident that the _______________ is between __________ and __________.

• A local elementary school wants to determine whether there is evidence to suggest that the boys and girls at the school play, on average, the same amount of time outside per week. A random sample of children from the school is surveyed regarding the number of hours they play outside per week. The statistic of interest is the difference (boys – girls) in the average number of hours per week boys and girls play outside. A 95% confidence interval to estimate the parameter of interest is (−1.9990, −0.4010).
   a) Complete the following statement:
      We are ____% confident that the _______________ is between __________ and __________.
   b) Based on the given confidence interval, what can we say about the claim that the average amount of time boys and girls from the school spend each week is the same? Explain.

• In Section 2.2 we considered the following scenario:

   According to their maker, the different colors of Skittles candies (red, orange, yellow, green, and purple) are produced in equal quantities. However, they are packaged by weight and size, not color distribution. Your friend does not believe they are produced equally; rather, a larger proportion of yellow Skittles are produced and packaged. To test this claim you order 10 of the “Fun Size” bags of Skittles from different distribution centers. You then empty all the bags into one bowl and count the number of yellow Skittles. You find that 16 of the 60 Skittles are yellow.

   Now let us construct a 95% confidence interval to estimate the population proportion of yellow Skittles produced. If the hypothesis \( H_0: \pi = 0.37 \) results in a p-value of 0.075 and the hypothesis \( H_0: \pi = 0.38 \) results in a p-value of 0.049, determine the 95% confidence interval for the population proportion of yellow Skittles produced and packaged. Explain your reasoning using complete sentences. Interpret the interval in context.

   Hint: See Example 4.1.2 for assistance.
Section 4.2 | Theory-Based Approach to Confidence Intervals

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fluently use the following vocabulary and symbols</td>
</tr>
<tr>
<td>○ Interval of plausible values</td>
</tr>
<tr>
<td>○ Confidence interval</td>
</tr>
<tr>
<td>○ Confidence level</td>
</tr>
<tr>
<td>○ Margin of error</td>
</tr>
<tr>
<td>• Use the 2SD Method to find an approximate 95% confidence interval for $\pi$ or $\mu$.</td>
</tr>
<tr>
<td>• Use the Theory-based Inference applet to find a confidence interval.</td>
</tr>
<tr>
<td>• Write an interpretation, in context, to give the real-world meaning of a confidence interval.</td>
</tr>
</tbody>
</table>
Methods for Constructing a Confidence Interval

1) **Multiple Tests of Plausibility Method**: This method, illustrated in the previous section, tests multiple null hypothesis values to determine which are plausible values for the parameter. Each plausible value is included in the confidence interval until we reject a value both above and below the observed statistic.

2) **2 Standard Deviations Method (2SD Method)**: For a 95% confidence level, we can estimate the confidence interval using the following formula:

   \[
   \text{(Observed Statistic) } \pm 2 \times \text{ (standard error)}
   \]

   Remember, the standard error refers to the standard deviation of the sampling distribution. The standard error can be approximated using the appropriate formula below.

   \[
   \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ for a categorical variable} \quad \text{or} \quad \frac{s}{\sqrt{n}} \text{ for a quantitative variable}
   \]

3) **Confidence Interval using an applet**: We can use the Theory-based Inference applet to find a confidence interval.

---

**Caution – Validity Conditions for a Confidence Interval for \( \pi \)**

When a sample of categorical data satisfies the conditions

\[\text{Number of Successes} \geq 10 \text{ and } \text{Number of Failures} \geq 10\]

we can use a THEORETICAL normal distribution to model the distribution of sample statistics (the distribution of \( \hat{p} \) values). If these conditions are not met, we should not use the theory-based approach and should instead use simulation.

**Caution – Validity Conditions for a Confidence Interval for \( \mu \)**

When a sample of quantitative data satisfies the conditions

\[\text{sample size is 20 or more} \text{ and } \text{sample data not strongly skewed}\]

we can use a THEORETICAL t-distribution to model the distribution of sample statistics (the distribution of \( \bar{x} \) values). If these conditions are not met, we should not use the theory-based approach and should instead use simulation.
Two types of medication for hives are being tested to determine whether there is a difference in their effectiveness. Twenty out of a random sample of 200 adults given Medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given Medication B still had hives 30 minutes after taking the medication.

(a) Create a 95% confidence interval to find a range of plausible values for the difference in the population proportion of people who are not helped by these medications (i.e., who still have hives 30 minutes after taking the medication).

(b) Interpret this interval in context.

Solution 4.2.1

(a) Let A and B be the subscripts for Medication A and Medication B, respectively. Then \( \pi_A \) and \( \pi_B \) are the unknown population proportions of people who are not helped by each respective medication.

We are given the observed statistic for each sample group, \( \hat{p}_A = \frac{20}{200} = 0.10 \), and \( \hat{p}_B = \frac{12}{200} = 0.06 \).

We calculate the observed statistic for the difference in proportions, \( \hat{p}_A - \hat{p}_B = 0.04 \).

We will create our confidence interval using the Theory-based Inference applet.

We start by changing the scenario to “Two proportions” and filling in the summary information for each group. Then check the “Confidence Interval” box and click the “Calculate CI” button.

The 95% confidence interval is (−0.013, 0.093).

(b) We are 95% confident that the difference in the proportion of all patients with hives 30 minutes after taking the medication A versus medication B is between -1.3% and 9.3%. We are 95% confident that the proportion of all adults who still have hives 30 minutes after taking Medication A is between 1.3 percentage points lower and 9.3 percentage points higher than the proportion of all adults who still have hives 30 minutes after taking Medication B.

Note that the value 0 (from the null hypothesis, indicating no difference in population proportions) is contained in this confidence interval. The confidence interval leads us to conclude that it is plausible that these two medications are equally effective at treating hives.
Statistics students want to know the mean score on a common final exam that was given in all sections of a statistics class at a large university. The overall exam statistics won’t be shared with the campus community for several months, but each student already received their own personal score on the exam (as a percentage out of 100). An interested group of students select a sample of ten statistics students and calculate a mean score of 67 with a standard deviation of 3.2.

(a) Create and interpret a 95% confidence interval for the average exam score.

(b) The bookstore promised discounts to all students who scored higher than the mean on this final exam. Three students want to know if they should expect to receive the discount at the bookstore. Their scores are as follows: Tim earned a 62, Tameka earned a 71, and Elenid earned a 66.

**Solution 4.2.2**

(a) We can use the 2SD Method to find an approximate 95% confidence interval for the average score on the final exam.

The midpoint of the confidence interval is the observed statistic, $\bar{x} = 67$.

The margin of error is calculated using the formula $\frac{s}{\sqrt{n}}$.

\[
\frac{s}{\sqrt{n}} = \frac{3.2}{\sqrt{10}} = 1.01
\]

We can then construct the confidence interval as follows:

\[
(\text{observed statistic}) \pm 2 \times \frac{s}{\sqrt{n}}
\]

\[
67 \pm 2 \times \frac{3.2}{\sqrt{10}}
\]

\[
67 \pm 2.02
\]

We can rewrite this as (64.98, 69.02), since $67 - 2.02 = 64.98$ and $67 + 2.02 = 69.02$.

We can interpret the interval as follows: We are 95% confident that the mean final exam score for all students is between 64.98 and 69.02.

(b) To be confident that they will get the discount at the bookstore, the students want to know if they did better than the mean. Let’s start with Tim, who scored a 62 on the exam. Since we are 95% confident that the mean final exam score is between 64.98 and 69.02, every plausible value is greater than Tim’s score. Tim is confident that he will not receive the discount.

Elenid scored a 66. Looking at the confidence interval, it is plausible that she scored higher than the mean. For example, 65 is inside the interval and if this is the true mean, Elenid gets the discount. On the other hand, the interval also includes plausible values that are greater than Elenid’s score. For example, 67 is inside the interval and if this is the true mean, then Elenid would not get the discount. Elenid cannot be confident either way. She’ll have to wait until the final exam statistics are released by the university to know whether she gets the discount.

Tameka’s score of 71 is greater than every plausible value in the confidence interval. She is confident that she will receive the discount.
We are interested in the average body temperature of human adults, which is commonly believed to be 98.6° F. To test whether this is true, we took a sample of 1000 people and found an average body temperature of 97.5° F with a standard deviation of 1° F.

(a) Create a 99% confidence interval to find a range of plausible values for the average body temperature of all human adults.
(b) Interpret this interval in context.
(c) The accepted ‘normal’ body temperature is 98.6° F. Based on the interval from part (b), is this plausible?

**Solution 4.2.3**

(a) Since the sample size is very large we can use a theory-based approach and we don’t have to worry about skewness in the sample data. We can find the confidence interval using the Theory-based Inference applet. We start by changing the scenario to “One mean” and filling in the summary information from the sample. Then check the “Confidence Interval” box, enter 99 in the “confidence level” box, and click the “Calculate CI” button.

The 99% confidence interval is (97.42, 97.58).

(b) We are 99% confident that the average body temperature of all adult humans is between 97.42° F and 97.58° F.

(c) Since 98.6° F is larger than every plausible value in the confidence interval, we can conclude that 98.6° F is not a plausible value for the average body temperature of all human adults.

For more information on this see: [Source](#).
**Example 4.2.4**

A business is interested in determining the percentage of customers, in the long run, who will make a last-minute purchase from the items located near the register. They collect a random sample of 85 customers and find that 12 of them made last-minute purchases.

(a) Use the 2SD method to create a 95% confidence interval for the percent of all customers, in the long run, who will make last-minute-purchases.

(b) Interpret this interval in context.

**Solution 4.2.4**

(a) We can use the 2SD Method to find an approximate 95% confidence interval. The midpoint of the confidence interval is the observed statistic, \( \hat{p} = \frac{12}{85} \approx 0.141 \) (14.1% of the sampled customers made last-minute purchases).

The margin of error is calculated using the formula \( \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \).

\[
\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{0.141(1 - 0.141)}{85}} = 0.038
\]

We can then construct the confidence interval as follows:

\[
(\text{observed statistic}) \pm 2 \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

\[
0.141 \pm 2 \times 0.038
\]

\[
0.141 \pm 0.076
\]

We can rewrite this as (0.065, 0.217) since 0.141 – 0.076 = 0.065 and 0.141 + 0.076 = 0.217.

(b) We can interpret the interval as follows: We are 95% confident that, in the long run, between 6.5% and 21.7% of all customers will make last-minute-purchases from items located near the register.
Section 4.2 | Theory-Based Approach to Confidence Intervals

Exercises

1) We want to estimate the average GPA of all students at American colleges. We take a random sample of 500 students and find an average GPA of 3.15 with a standard deviation of 0.275.
   a) Use the 2SD Method to create a 95% confidence interval for the average GPA of American college students. Show all your work.
   b) Complete the following statement:
      \[ \text{We are } \% \text{ confident that the average GPA is between } \text{ and } \text{.} \]
   c) In the past, the “average” grade was considered to be a C, which corresponds to a GPA of 2. Based on the confidence interval, do we believe that the average grade is now higher than a C? Explain.

2) You read online that the probability that a person picked at random in Times Square in New York City is a tourist visiting the area is 0.83. You are interested in this so you take a sample of 100 random people in Times Square and find that 75% of the sample are tourists.
   a) Use the 2SD Method to create a 95% confidence interval to estimate the probability that a person picked at random in Times Square in New York City is a tourist visiting the area. Show all your work.
   b) Complete the following statement:
      \[ \text{We are } \% \text{ confident that the probability is between } \text{ and } \text{.} \]
   c) Based on the confidence interval, is it plausible that the online source is correct? Explain.

3) It has been claimed that in competitive sports there might be some advantage for teams wearing blue uniforms. You want to test this with the sport of competitive swimming. You conduct a study using 36 separate races where 3 competitive swimmers are chosen at random and assigned either a blue, red, or green swimsuit for each race. The swimmer wearing a blue swimsuit won in 18 of the 36 races.
   a) Use the 2SD Method to calculate a 95% confidence interval for the proportion of all races that are won by a swimmer wearing a blue swimsuit. Show all your work and round proportions to 4 decimal places.
   b) Interpret the confidence interval in context.

4) In 1948 the Chicago Daily Tribune erroneously published a headline due to early polling results that indicated Thomas E. Dewey had defeated Harry S. Truman in that year’s presidential election. In truth, Truman won 49.6% of the popular vote (24,179,347 out of 47,346,569 votes) to Dewey’s 45.1% (21,991,292 out of 47,346,569 votes), with the remaining percentage going to third parties.
   The Gallup’s September 24 report indicated that they had obtained a random sample of 3,250 voters, and 46.5% of sampled voters indicated they would be voting for Dewey, while only 38% planned to vote for Truman.
   a) Use the 2SD Method to calculate a 95% confidence interval for each parameter. Show all your work and round proportions to 4 decimal places.
   b) Interpret each of the confidence intervals in context.

5) The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years. After learning this you decide to collect data from a random sample of New York City obituaries from 2018-2020 to compare the life expectancy of New York City residents to all United States citizens. After surveying 42 obituaries, you find a mean life expectancy of 81.9 years with a sample standard deviation of 8.3 years.
   a) Use the 2SD Method to calculate a 95% confidence interval for the parameter of interest. Show all your work and round values to 4 decimal places.
   b) Interpret the confidence interval in context.
   c) Use the confidence interval to decide whether it is plausible that the average life expectancy of New York City residents is the same as the national average life expectancy in 2018. Explain your decision.
6) A group of National Fish and Wildlife researchers are collecting data on the grizzly bear population in Bearenstein National Park. It is estimated that there are 640 grizzly bears in the park. In previous studies the researchers found evidence that a parasite was present in a significant portion of the population of grizzly bears in the park, the parasite is causing significant weight change in the male grizzly population, and a higher proportion of males are infected with the parasite than females.

Now, researchers wish to estimate the population proportion of male bears infected with the parasite. The researchers collected a new random sample of 105 bears at Bearenstein National Park. Summary results from the study are shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parasite</td>
<td>22</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>No Parasite</td>
<td>83</td>
<td>32</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>47</td>
<td>58</td>
</tr>
</tbody>
</table>

(a) Use the Theory-based Inference applet to create a 99% confidence interval for the parameter of interest. Show all your work and round proportions to 4 decimal places.

(b) Use the confidence interval to fill in the blanks and complete the following statements:

I am ______ % confidence that the proportion of all male grizzly bears in Bearenstein National Park infected with the parasite is between _____________ and _____________.

I am ______ % confident that the number of male grizzly bears in Bearenstein National Park infected with the parasite is between _____________ and _____________. bears.

7) Kelley Blue Book is a vehicle valuation and automotive research company that reports market value prices for new and used automobiles. These values are often used by auto dealerships to determine the selling prices of their vehicles. The average Kelley Blue Book value of a used 2008-2012 sedan is $5,560.

I would like to determine whether I can ask for a different selling price in my region for my sedan, so I conduct a random sample of local ads for used 2008-2012 sedans and record the asking price for 41 of them. I calculate an average asking price of $5,985 with a standard deviation of $765.

(a) Use the Theory-based Inference applet to create a 90% confidence interval for the parameter of interest. Show all your work and round values to 4 decimal places.

(b) Interpret the confidence interval in context.

(c) What should I choose for the asking price of my car according to your answer to part (b)? Explain.
Section 4.3 | Impacts on the Width of a Confidence Interval

<table>
<thead>
<tr>
<th>Section Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Understand how changing the sample size will affect the width of a confidence interval.</td>
</tr>
<tr>
<td>• Understand how changing the confidence level will affect the width of a confidence interval.</td>
</tr>
</tbody>
</table>

Confidence Level - Reminder

The most commonly used confidence levels and their corresponding significance levels are given in the following table. Recall that, when writing the confidence level in decimal form, \((\text{Significance Level})+(\text{Confidence Level})=1\).

<table>
<thead>
<tr>
<th>Significance Level, (\alpha)</th>
<th>Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>0.10</td>
<td>0.90</td>
</tr>
</tbody>
</table>

To understand how changing the confidence level will impact the width of the confidence interval, we need to dig into the connection to the p-value a little more.

Let’s consider two significance levels, \(\alpha = 0.05\) and \(\alpha = 0.01\):

\[\alpha = 0.05\]

If the p-value is less than or equal to 0.05 then we have sufficient evidence to reject the null hypothesis and conclude that the null hypothesis is not plausible. If the p-value is greater than 0.05 then we conclude that the null hypothesis is plausible.

\[\alpha = 0.01\]

If the p-value is less than or equal to 0.01 then we have sufficient evidence to reject the null hypothesis and conclude that the null hypothesis is not plausible. If the p-value is greater than 0.01 then we conclude that the null hypothesis is plausible.

Notice the green dots on the number lines above. If these dots represent the p-value obtained from testing a claim about the value of the parameter (notice that it is the same p-value in both cases), then we would reject the null hypothesis for \(\alpha = 0.05\) but conclude that the null hypothesis is plausible for \(\alpha = 0.01\).
A confidence interval is the set of all plausible values for the parameter. That is, it is the set of every possible parameter value for which our p-value lands in the purple region of the number line. Since the purple region is larger for $\alpha = 0.01$, more parameter values will turn out to be plausible when $\alpha = 0.01$. A smaller significance level, $\alpha$, leads to a wider confidence interval, and a smaller significance level corresponds to a larger confidence level.

- A smaller significance level, $\alpha$, results in a wider confidence interval.
- A larger confidence level results in a wider confidence interval.

This agrees with the idea that making the significance level smaller will make it harder to reject the null hypothesis. The harder it is to reject a possible parameter value, the more of them you will conclude are plausible. To summarize, holding all other things constant, increasing the confidence level will increase the width of the interval!

**LOGIC TO UNDERSTAND THIS:**

It makes logical sense to say that we will have to cast a “wider net” around the observed statistic if we want to have more confidence in our conclusion. For example, we could say that we are 100% confident that the true value of the population parameter is somewhere between $-\infty$ and $\infty$. Thus, having a very high confidence level would correspond to having a very wide interval. But, if we shrink down that confidence interval to a small interval surrounding our observed statistic, then we will be less confident that we’ve actually captured the true value of the population parameter within that small interval.

**Example 4.3.1**

A production line produces boxes of cereal. The line supervisor wants to know the average ounces of cereal being put into the boxes. The supervisor collects a random sample of boxes and creates the following three confidence intervals:

- 90%: (20.07, 21.93)
- 95%: (19.88, 22.12)
- 99%: (19.50, 22.50)

(a) Interpret each of the intervals in context.

(b) What is the effect of changing the confidence level on the width of the interval?

(c) The production line is supposed to be filling the boxes with 20 ounces on average. Do the confidence intervals indicate that the boxes are being filled with the right amount of cereal?

**Solution 4.3.1**

(a) Interpretations:

- 90%: We are 90% confident that the average amount of cereal in all boxes is between 20.07 and 21.93 ounces.
- 95%: We are 95% confident that the average amount of cereal in all boxes is between 19.88 and 22.12 ounces.
- 99%: We are 99% confident that the average amount of cereal in all boxes is between 19.50 and 22.50 ounces.

(b) The narrowest interval is from the 90% confidence level with a width of 21.93 – 20.07 = 1.86 ounces. As the confidence level increases to 95% and 99%, the width increases to 2.24 ounces and 3 ounces, respectively.

(c) The first confidence interval (90% confidence level) has numbers all larger than 20 ounces. Based only on this interval, we may think that the production line is over filling the boxes. As we increase the confidence level, we see that we get intervals that do include 20 ounces. Based on the last two intervals (the 95% and the 99% intervals), we see that 20 ounces is a plausible value for the average amount of cereal in all boxes of cereal.
**Sample Size**

We also need to consider the effect of sample size on the width of a confidence interval. Remember that the larger the sample size, the stronger the evidence against the null hypothesis. Increasing the sample size causes the distribution of sample statistics to become narrower because bigger sample sizes have distributions of sample statistics that have smaller standard deviations. Thus, the same statistic becomes further out in the tail of the narrower distribution. A more extreme (further in the tail) statistic results in a smaller p-value. This means that the same hypothesized parameter value could be plausible with one sample size, but rejected with a larger sample size!

---

**Example 4.3.2**

Politicians want to understand the voting population’s support for a new tax.
- Basheer takes a random sample of 150 registered voters and finds that 45 of them support the new tax.
  \[
  \hat{p}_{\text{Basheer}} = \frac{45}{150} = 0.30 = 30\%.
  \]
- Fatima takes a random sample of 1500 registered voters and finds that 450 of them support the new tax.
  \[
  \hat{p}_{\text{Fatima}} = \frac{450}{1500} = 0.30 = 30\%.
  \]

Suppose Basheer and Fatima each construct a 95% confidence interval based on their sample results. Whose confidence interval will be wider?

**Solution 4.3.2**

Before we do the calculation, let us try and think through what will happen. We notice that both Basheer and Fatima have the same observed statistic, 30%. The difference between them is the sample size, 150 versus 1500. As we saw in Example 2.4.2, Fatima’s sample statistic provides stronger evidence against the null hypothesis, which makes sense because a bigger sample would provide stronger evidence. This was shown by looking at both Fatima’s and Basheer’s distribution with the assumption \( \pi = 0.35 \) (as shown again on the next page).

The key point here is that this will be true for every possible null hypothesis value, not just the single value we tested in Example 2.4.2. Each possible parameter value in the null hypothesis will lead to a smaller p-value for Fatima than for Basheer, which will lead to more of Fatima’s hypothesized parameter values being rejected. More rejected values means a smaller confidence interval. Fatima will have more rejected parameter values, and thus a smaller interval. Basheer will have the wider interval!

**Intervals:** We can find the confidence intervals using the **Theory-based Inference** applet.

- Basheer: (0.2267, 0.3733) or approximately 23% to 37%. This confidence interval has a width of 14 percentage points.
- Fatima: (0.2768, 0.3232) or approximately 28% to 32%. This confidence interval has a width of 4 percentage points.

Basheer has the wider interval of plausible values because he used a smaller sample size.
The simulated distributions for both Basheer and Fatima are given below using the same scale along the horizontal axis.

To summarize, holding all other things constant, the larger the sample size, the smaller the confidence interval.
Section 4.3 | Impacts on the Width of a Confidence Interval

Exercises

1) A sample of 20 heads of lettuce is selected from a crop. The weight of each head of lettuce is recorded and the mean weight is 2.2 pounds with a standard deviation of 0.1 pounds.

   (a) Use the Theory-based Inference applet to create a 95% confidence interval for the average weight of a head of lettuce from this crop. Interpret the interval in context.
   (b) Use the Theory-based Inference applet to create a 99% confidence interval for the average weight of a head of lettuce from this crop. What aspects of the interval from (a) stayed the same? What aspects changed?
   (c) Suppose instead that the standard deviation of the sample was 0.2 pounds. Use this new standard deviation to create a 95% confidence interval for the average weight of a head of lettuce from this crop. What aspects of the interval from part (a) stayed the same? What aspects changed?

2) A random sample of 1,200 voters were asked what the most significant issue was in the upcoming election. Sixty-five percent answered, “the economy”.

   (a) Use the Theory-based Inference applet to create a 90% confidence interval for the proportion of all voters who believe the economy is the most significant issue in the upcoming election. Interpret the interval in context.
   (b) Suppose that 600 voters were sampled and sixty-five percent answered, “the economy”. What aspects of the interval from part (a) would stay the same? What aspects would change? Check your answers by finding the 90% confidence interval for this scenario.

3) A study was conducted by the National Endowment for the Arts to described music preferences in the Hudson Valley. The 2010 report describes relationships between key demographic characteristics and music preferences. It found that 24% of adults preferred the genre of Rock/Heavy Metal at that time. Recently, a random sample of 180 Hudson Valley adults was surveyed with 56 indicating they prefer Rock/Heavy Metal.

   (a) Use the 2SD Method to create a 95% confidence interval for the proportion of all Hudson Valley adults who prefer Rock/Heavy Metal. Interpret the interval in context.
   (b) Suppose that a random sample of 180 Hudson Valley adults found that 57 of them prefer Rock/Heavy Metal. What aspects of the interval from part (a) would stay the same? What aspects would change? Check your answers by finding the 95% confidence interval for this scenario.

4) The Centers for Disease Control (CDC) determined that the average life expectancy for the U.S. population in 2018 was 78.7 years with a standard deviation of 9.4 years. After learning this you decide to collect data from a random sample of New York City obituaries from 2018 to 2020 to compare the life expectancy of New York City residents to all United States citizens. After surveying 75 obituaries, you find a mean life expectancy of 81.7 years with a standard deviation of 8.3 years.

   (a) Use the 2SD Method to create a 95% confidence interval for the mean life expectancy of all New York City residents during 2018 to 2020. Interpret the interval in context.
   (b) Suppose that the random sample had a standard deviation of 13.1 years instead of 8.3 years. What aspects of the interval from part (a) would stay the same? What aspects would change? Check your answers by finding the 95% confidence interval for this scenario.
5) The Pepsi Challenge is an ongoing marketing campaign run by PepsiCo since 1975 in which participants complete a blind taste test of two sodas (Pepsi and their primary competitor, Coca-Cola) and select their preferred soda. PepsiCo claims that people are more likely to prefer their brand of cola to Coca-Cola. A random sample of 351 people is given the taste test, and 54% of them choose Pepsi as their preferred cola.

(a) Use the 2SD Method to create a 95% confidence interval for the proportion of all people who would choose Pepsi as their preferred cola. Interpret the interval in context.
(b) Suppose that a random sample of 513 people participated in the challenge, and 54% of them chose Pepsi as their preferred cola. What aspects of the interval from part (a) would stay the same? What aspects would change? Check your answers by finding the 95% confidence interval for this scenario.

6) A 2018 article by the New York Times investigating the average commute time of workers in major cities found that the longest was for workers who live in New York City, who traveled an average of 35.9 minutes each day with a standard deviation of 12.5 minutes. You live relatively close to the city and know that many of your neighbors work there, so you decide to investigate whether they have a longer average commute time than the average worker who lives in New York City. You survey 38 commuters in your neighborhood and find an average commute time of 39.7 minutes with a standard deviation of 6.1 minutes.

(a) Use the Theory-based Inference applet to create a 90% confidence interval for the mean commute time for all commuters in your neighborhood. Interpret the interval in context.
(b) Suppose that we create a 99% confidence interval using the same sample data. What aspects of the interval from part (a) would stay the same? What aspects would change? Check your answers by finding the 99% confidence interval.
Chapter 5 | LINEAR REGRESSION AND CORRELATION

Contents of Chapter 5

Section 5.1 Scatterplots and Review of Linear Equations

Section 5.2 The Regression Equation

Section 5.3 Prediction and Residuals

Section 5.4 Testing for Correlation

Introduction

Professionals often want to know how two or more quantitative variables are related. For example, is there a relationship between student’s grades on the second math exam and student’s grades on the final exam? If there is a relationship, what is the relationship and how strong is it? This type of data (where we collect both the final exam score as well as the second exam score) is bivariate data — "bi" for two variables. In many real-world situations, statisticians use multivariate data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression," with one independent variable, often denoted by x. This involves data that is reasonably well described by a line. You will also study correlation, which measures how strong the linear relationship is between the two variables.
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relationship between two quantitative variables, $x$ and $y$. The most common way is a scatter plot. The following example illustrates a scatter plot.

**Example 5.1.1**

In Europe and Asia, M-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets and also provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet.

The data below gives the number of M-commerce users for certain years. Use the data to construct a scatter plot. Let $x$ = the year and let $y$ = the number of M-commerce users, in millions.

<table>
<thead>
<tr>
<th>$x$ (year)</th>
<th>$y$ (# of users in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.5</td>
</tr>
<tr>
<td>2002</td>
<td>20.0</td>
</tr>
<tr>
<td>2003</td>
<td>33.0</td>
</tr>
<tr>
<td>2004</td>
<td>47.0</td>
</tr>
</tbody>
</table>

The data points from the table can be interpreted in context as follows:

(2000, 0.5): “In the year 2000, there were 500,000 M-commerce users”.
(2002, 20.0): “In the year 2002, there were 20 million (20,000,000) M-commerce users”.
(2003, 33.0): “In the year 2003, there were 33 million (33,000,000) M-commerce users”.
(2004, 47.0): “In the year 2004, there were 47 million (47,000,000) M-commerce users”.
Using an Applet to Create a Scatterplot

To create a scatter plot, we can use the Two Quantitative Variables applet as explained below.

1. Paste the data points into the Two Quantitative Variables applet.

2. Note that column headings (footlength and height in the screen shot at right) cannot contain spaces, data values should be separated by a tab or space, and each ordered pair should be on its own line.

3. Click the "Use Data" button.
   Note: You can add, remove, or manipulate data points using the "Add/Remove Observations" and the "Move observations" options at the bottom of the applet.

Try It

5.1.1

Amelia plays basketball for her high school. She wants to improve so that she can play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

<table>
<thead>
<tr>
<th>x (hours practicing jump shot)</th>
<th>y (points scored in a game)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>12</td>
<td>36</td>
</tr>
</tbody>
</table>

Construct a scatter plot and state whether the data seem to indicate that Amelia tends to score more points in a game when she spends more time practicing her jump shot.
When considering a scatter plot, we often wish to describe the scatter plot in terms of **form**, **direction**, and **strength**.

The **form** of a scatter plot refers to the pattern that the points create. In general, we are interested in whether the points seem to form a **linear** versus a **non-linear** pattern. Note that the points do not need to fall in a perfect or nearly perfect straight line for us to conclude that the form of the scatter plot is linear. We will use the strength of the scatter plot to indicate how closely the points form a straight line pattern. We can also use the phrase **no relationship** to indicate that the points do not seem to form a pattern of any kind. This includes points that fall along a horizontal line or closely resemble a horizontal line pattern. In the scatter plots on the next page, the top right graph shows the points forming a definite non-linear pattern. The bottom right graph shows points that indicate no relationship (no pattern).

The **direction** of a scatter plot refers to whether the points seem to be increasing (going up) or decreasing (going down) as we read the graph from left to right. A clear direction is visible when either of the following situations occurs.

- High values of one variable occur with high values of the other variable, and low values of one variable occur with low values of the other variable. In this case the points are increasing and the **direction** of the scatter plot is **positive**.
- High values of one variable occur with low values of the other variable. In this case the points are decreasing and the **direction** of the scatter plot is **negative**.

The **strength** of a scatter plot refers to how closely the points resemble a straight line pattern.

- Scatter plots whose points very closely resemble a straight line pattern have a **strong** linear relationship.
- Scatter plots whose points somewhat resemble a straight line pattern have a **moderate** linear relationship.
- Scatter plots whose points only loosely resemble a straight line pattern have a **weak** linear relationship.

The scatter plots below illustrate different strengths and patterns.
Review of Linear Equations

Linear equations can be written in the form \( y = ax + b \) where

- \( x \) is often the letter used to denote the **independent variable**, also called the **input variable** or the **explanatory variable**.
- \( y \) is often the letter used to denote the **dependent variable**, also called the **output variable** or the **response variable**.
- The point \((0, a)\) is the vertical intercept.
- The number \( b \) is the **slope** of the line. Every time \( x \) increases by 1 unit, we add \( b \) to the \( y \)-value. If the slope is a positive number, then the graph will be increasing from left to right. If the slope is a negative number, then the graph will be decreasing from left to right.

Note that in algebra the equation of a line is often written \( y = mx + b \). In this form, the point \((0, b)\) is the vertical intercept and the number \( m \) is the slope of the line. In statistics, the number \( b \) is the slope, **not** the vertical intercept.

### Example 5.1.2

Following are examples of linear equations.

\[
y = 3 + 2x \quad \text{The vertical intercept is} \ (0,3) \ \text{and the slope of the line is} \ 2.
\]

\[
y = -0.01 + 1.2x \quad \text{The vertical intercept is} \ (0, -0.01) \ \text{and the slope of the line is} \ 1.2.
\]
5.1.2

Is the following an example of a linear equation?

\[ y = -0.125 - 3.5x \]

The graph of a linear equation of the form \( y = a + bx \) is a straight line. Any non-vertical line can be written in this form. As shown below, when graphing a linear equation the horizontal axis represents the explanatory (or independent or input) variable and the vertical axis represents the response (or dependent or output) variable.
Example 5.1.3

(a) Graph the equation \( y = -1 + 2x \).

(b) Suppose this equation represents the profit a company earns, \( y \), in millions of dollars, where \( x \) is the number of items sold in thousands. Write a complete sentence to interpret the contextual meaning of the point (3,5), which is a point on the graph of this line.

Solution 5.1.3

(a) Relating this equation to the generic form \( y = a + bx \), we see the following:

\[ a = -1 \] indicates that the graph will cross the vertical axis at the point (0, -1). This is the vertical intercept.

\[ b = 2 \] indicates the slope of the graph: every time the \( x \) value increases by 1 unit, the \( y \) value increases by 2 units.

For example, the graph passes through the point (0, -1) (the \( y \)-intercept)

and the point (1, 1) (start at previous point, travel to the right 1 unit and up 2 units)

and the point (2, 3) (start at previous point, travel to the right 1 unit and up 2 units)

and the point (3, 5) (start at previous point, travel to the right 1 unit and up 2 units)

When these points are connected with a straight line that extends in both directions, the graph of the equation is created.

(b) The point (3,5) indicates that when \( x = 3 \), then \( y = 5 \). In context, \( x \) represents the number of items sold in thousands, so \( x = 3 \) means 3000 items were sold. In context, \( y \) represents the profit in millions of dollars, so \( y = 5 \) this means that the profit is $5,000,000. The sentence that interprets the point (3,5) is as follows: “When the company sells 3000 items, the company makes a profit of $5,000,000.”
Example 5.1.4

Aaron’s Word Processing Service (AWPS) does word processing (editing documents for customers). The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job. Find the equation that expresses the total cost, in dollars, in terms of the number of hours required to complete the job.

Solution 5.1.4

Let \( x \) = the number of hours it takes to get the job done. Let \( y \) = the total cost to the customer in dollars.

The $31.50 is the fixed cost (the initial cost) because when \( x = 0 \) hours, then \( y = 31.50 \) dollars of total cost.

If it takes \( x \) hours to complete the job, then \((32 \text{ dollars/unit}) \times (x \text{ hours})\) is the cost of the word processing only.

The total cost is \( y = 31.50 + 32x \).

A graph of the total cost function is shown below. Notice that the time to complete the job in hours, the explanatory (or independent or input) variable, is depicted on the horizontal axis. The total cost in dollars, the response (or dependent or output) variable, is depicted on the vertical axis.
5.1.3
Compare the two equations and graphs from Examples 5.1.3 and 5.1.4. What do the values "32" and "31.5" represent in the equation from Example 5.1.4? What does the point (2, 95.5) mean in the context of Example 5.1.4?

5.1.4
Emma’s Extreme Sports hires hang-gliding instructors and, for each class taught, pays them a fee of $50 plus $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost, in dollars, in terms of the number of students in a class.
Slope and $y$-intercept of a Linear Equation

For the linear equation $y = a + bx$, $b$ represents the slope and $a$ represents the $y$-intercept. Recall from algebra that the slope is a number that describes the steepness and direction of a line, and the $y$-intercept, which has coordinates $(0, a)$, is the point where the line crosses the $y$-axis.

There are three possible situations with regard to the value of the slope, $b$. These are described and illustrated below.

(a) If $b > 0$, the line slopes upward to the right.

(b) If $b = 0$, the line is horizontal.

(c) If $b < 0$, the line slopes downward to the right.

Example 5.1.5

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

(a) Identify (describe in words) the independent and dependent variables, and state what letters are used to represent them in the equation.

(b) What are the values of the $y$-intercept and slope? Interpret these values in context with complete sentences.

Solution 5.1.5

(a) The independent variable, $x$, is the number of hours Svetlana tutors each session. The dependent variable, $y$, is the amount, in dollars, Svetlana earns for each tutoring session.

(b) The $y$-intercept is the point $(0, 25)$. In context, this can be interpreted as follows: “At the start of the tutoring session, Svetlana charges a one-time fee of $25”. The slope is 15. In context, this can be interpreted as follows: “Svetlana increases her earnings by $15 for each hour she tutors”.
5.1.5

Ethan repairs household appliances such as dishwashers and refrigerators. He charges $25 for each visit, plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is \( y = 25 + 20x \).

(a) Identify (describe in words) the independent and dependent variables, and state what letters are used to represent them in the equation.

(b) What are the values of the \( y \)-intercept and slope? Interpret these values in context with complete sentences.
Exercises

1) Answer each question below for each of the six graphs.

(a) Describe the direction (positive or negative) of the scatter plot, if a direction exists.
(b) Describe the form of the scatter plot (linear, not linear, no relationship).
(c) Describe the strength of the relationship (strong, moderate, weak).

2) Answer each question below for each of the six graphs.

(a) Describe the direction (positive or negative) of the scatter plot, if a direction exists.
(b) Describe the form of the scatter plot (linear, not linear, no relationship).
(c) Describe the strength of the relationship (strong, moderate, weak).
3) A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25, plus $12.50 per hour.
   (a) Identify (describe in words) the dependent and independent variables.
   (b) Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

4) A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.
   (a) Identify (describe in words) the dependent and independent variables.
   (b) Find the equation that expresses the total charge in dollars in terms of the number of days the payment is late.

5) Is the equation \( y = 10 + 5x - 3x^2 \) linear? Explain.

6) Which of the following equations are linear? Explain your reasoning.
   - \( y = 6x + 8 \)
   - \( y + 7 = 3x \)
   - \( y - x = 8x^2 \)
   - \( 4y = 8 \)

7) Does the graph at right show a linear equation? Explain.

8) If we have two variables, "year" and "number of flu cases diagnosed," explain which variable makes the most sense as the independent variable and which makes the most sense as the dependent variable.

9) A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee in dollars that the company charges for each session is \( y = 50 + 100x \).
   (a) Identify (describe in words) the independent and dependent variables, and state what letters are used to represent them in the equation.
   (b) What are the coordinates of the \( y \)-intercept? Write a sentence to interpret the \( y \)-intercept in context.
   (c) What is the value of the slope? Write a sentence to interpret the slope in context.
10) Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil in pounds lost per year is \( y = 12,000x \).

(a) Identify (describe in words) the explanatory and response variables, and state what letters are used to represent them in the equation.

(b) What are the coordinates of the \( y \)-intercept? Write a sentence to interpret the \( y \)-intercept in context.

(c) What is the value of the slope? Write a sentence to interpret the slope in context.

11) The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the share price for Shipment Express is \( y = 15 – 1.5x \), where \( x \) is the number of hours passed since the opening of the trading day.

(a) What are the coordinates of the \( y \)-intercept? Write a sentence to interpret the \( y \)-intercept in context.

(b) What is the value of the slope? Interpret the meaning of the slope in context.

(c) If you owned this stock, would you want the price equation to have a positive or negative slope? Explain.

12) For each of the following situations, identify (describe in words) the independent and dependent variables.

- A study is done to determine whether elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
- A study is done to determine whether the weekly grocery bill changes based on the number of family members.
- Insurance companies base life insurance premiums partially on the age of the applicant.
- Utility bills vary according to power consumption.
- A study is done to determine whether, as the proportion of citizens with a college degree increases, the crime rate decreases.
5.2 | The Regression Equation

Data in the real world rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to be reasonably well represented by a straight line. This is called a Line of Best Fit or Least Squares Regression Line.

In this chapter, we are interested in scatter plots that show a linear pattern. The linear relationship is strong if the points are close to a straight line (except in the case of a horizontal line where there is no relationship). If the points show a linear relationship, then we would like to draw a line on the scatter plot that best models that linear relationship. This line can be calculated through a process called linear regression. However, we only calculate a regression line if values of one of the variables helps to explain or predict values of the other variable (i.e., if we believe that the variables are correlated). If \( x \) represents the independent variable and \( y \) represents the dependent variable, then we can use a regression line equation to predict the value of \( y \) for a given value of \( x \).

### Collaborative Exercise

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? In groups of about 6 students, collect data on both pinky finger length, in inches, and height in inches from each person in the group. The independent variable, \( x \), is pinky finger length in inches and the dependent variable, \( y \), is height in inches. Plot the points on graph paper (a blank grid is provided below). As a group, come to agreement on the appearance of the line that would best fit the points, and then **use a ruler** to "by eye" draw a line that appears to "fit" the data. Pick two convenient points on the line you drew and use them to find the slope of the line. Find the \( y \)-intercept of the line by extending your line so it crosses the \( y \)-axis. Using the slope and the \( y \)-intercept, write your equation of "best fit". Do you think all groups in the class will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?
Example 5.2.1

A random sample of 11 statistics students produced the following data, where \( x \) represents the unit 3 exam score out of 80 total points, and \( y \) represents the final exam score out of 200 total points. Can you predict the final exam score of a student if you know her/his third exam score?

<table>
<thead>
<tr>
<th>( x ) (third exam score)</th>
<th>( y ) (final exam score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>175</td>
</tr>
<tr>
<td>67</td>
<td>133</td>
</tr>
<tr>
<td>71</td>
<td>185</td>
</tr>
<tr>
<td>71</td>
<td>163</td>
</tr>
<tr>
<td>66</td>
<td>126</td>
</tr>
<tr>
<td>75</td>
<td>198</td>
</tr>
<tr>
<td>67</td>
<td>153</td>
</tr>
<tr>
<td>70</td>
<td>163</td>
</tr>
<tr>
<td>71</td>
<td>159</td>
</tr>
<tr>
<td>69</td>
<td>151</td>
</tr>
<tr>
<td>69</td>
<td>159</td>
</tr>
</tbody>
</table>

Solution 5.2.1

The third exam score, represented by \( x \), is the independent variable and the final exam score, represented by \( y \), is the dependent variable. We will plot a regression line that best "fits" the data. If each student in the class were to fit a line "by eye," they would most likely draw different lines. We can use what is called a least squares regression line to obtain the best fit line.

Consider the diagram below. Each point of observed data is of the form \((x, y)\) and each point on the best-fit line has the form \((x, \hat{y})\). The \( \hat{y} \) is read "\( y \) hat" and is the estimated value of \( y \) for a particular value of \( x \). It is the value of \( y \) obtained using the regression line equation, and may not equal the value of \( y \) from the observed data.

(continued on next page)
Solution 5.2.1 (continued):

The term \( y - \hat{y} = \varepsilon \) is called the "error" or “residual”. It is not an error in the sense of a mistake. It is the difference between the y-value of the observed data point and the corresponding predicted \( \hat{y} \)-value from the regression line of best fit.

The **absolute value of a residual** measures the vertical distance between the observed value of \( y \) and the estimated value of \( y \). In other words, it measures the vertical distance between the observed \( y \)-value of the data point and the predicted \( y \)-value from the point on the regression line.

- If the observed data point lies above the line, then the residual is positive and the best-fit line underestimates the observed value for \( y \).
- If the observed data point lies below the line, then the residual is negative and the best-fit line overestimates the observed value for \( y \).

In the diagram above, \( y_0 - \hat{y}_0 = \varepsilon_0 \) is the residual for the point \((x_0, y_0)\). For the given \( x \)-value \((x_0)\), the observed \( y \)-value \((y_0)\) lies above the \( y \)-value on the regression line \((\hat{y}_0)\), and so the residual is positive.

The least squares regression line (line of best fit) for the third exam/final exam scenario has the equation \( \hat{y} = -173.51 + 4.83x \). Consider the student who scored 65 (out of 80) on the third exam and 175 (out of 200) on the final exam (the first entry in the table). The regression equation predicts that this student will have a final exam score of 140.44: \( \hat{y} = -173.51 + 4.83(65) = -173.51 + 313.95 = 140.44 \). The residual in this case is \( 175 - 140.44 = 34.56 \). That is, the regression equation underestimates this student’s final exam score by about 35 points.

**Which Line is the Best Fit Line?**

For each observed data point in Example 5.2.1, we can calculate the residual: \( y_i - \hat{y}_i = \varepsilon_i \) for \( i = 1, 2, 3, ..., 11 \).

The magnitude (absolute value) of each \( \varepsilon_i \) is the vertical distance between the \( y \)-value of the observed data point and the corresponding \( y \)-value on the regression equation. In Example 5.2.1, there are 11 data points. Therefore, there are 11 \( \varepsilon \) values.
If you square each \( \varepsilon \) and then add the squared values together, the result is called the **Sum of Squared Errors (SSE)**.

The line of best fit, also called the regression equation, is the **line which minimizes the Sum of Squared Errors (SSE)**. That is, the “least squares” regression line is the line that makes the SSE smallest.

### Least Squares Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line: that the data are linearly correlated. The criteria for the best-fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line (with different values for the slope and vertical intercept) would have a higher SSE than the best-fit line. This best fit line is called the **least squares regression line**.

**NOTE**

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line. The calculations tend to be tedious if done by hand, so technology is used to find regression equations.

We will use the **Two Quantitative Variables** applet to find the equation for the least squares regression line, as described on the next page.
Using an Applet to Find the Regression Equation

**Use applet to create a scatter plot** (this was explained in Section 5.1)

1. **Paste the data points into the Two Quantitative Variables applet.**
   *For example, the data to the right has x-values representing a person’s foot length in cm, and y-values representing the person’s height in inches.*

2. **Note that column headings (which are used as axis labels) cannot contain spaces, data values should be separated by a tab or space, and each ordered pair should be on its own line.**

3. **Click the "Use Data" button.**
   
   *Note: You can add, remove, or manipulate data points using the "Add/Remove Observations" and the "Move observations" options at the bottom of the applet.*

**Use applet to identify and display the least squares regression equation:**

1. **Take a moment to estimate where the line should pass through to minimize the residuals.**

2. **After the data has been displayed in the scatter plot, check the "Show Regression Line" box to display the least squares regression equation and the line on the scatter plot.**

   In the screen shot above, we see that the regression equation is \( \text{height} = 38.30 + 1.03 \times \text{foot length} \), which indicates that in order to estimate a person’s height in inches, you would multiply foot length by 1.03 and then add 38.30. Note that \( \text{height} \) is read “predicted height”. The regression (best-fit) equation is \( \hat{y} = 38.30 + 1.03x \), where \( x \) represents foot length in cm and \( \hat{y} \) represents the predicted height in inches.
UNDERSTANDING SLOPE OF A REGRESSION LINE

It is important to interpret the slope of the regression line in the context of the situation represented by the data. You should be able to write a sentence interpreting the value of the slope in everyday understandable language.

INTERPRETATION OF THE SLOPE

The slope of the best-fit line tells us how much the dependent variable, $y$, is predicted to change for every one unit increase in the independent variable, $x$, on average. Positive slope represents an increasing line, and negative slope represents a decreasing line.

EXAMPLE 5.2.1 REVISITED

For the data from Example 5.2.1 regarding student exam scores, the slope of the regression line is $b = 4.83 = 4.83/1$.

The units of slope are (units of output)/(units of input) = (points on final exam)/(points on exam 3)

We can interpret the slope as follows: For every one-point increase in the third exam score, the final exam score is predicted to increase by 4.83 points on average.

### Try It

5.2.1

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below show different depths in feet with corresponding maximum dive times in minutes. Use the Two Quantitative Variables applet to find the best-fit equation.

<table>
<thead>
<tr>
<th>$x$ (depth in feet)</th>
<th>$y$ (maximum dive time in minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>60</td>
<td>55</td>
</tr>
<tr>
<td>70</td>
<td>45</td>
</tr>
<tr>
<td>80</td>
<td>35</td>
</tr>
<tr>
<td>90</td>
<td>25</td>
</tr>
<tr>
<td>100</td>
<td>22</td>
</tr>
</tbody>
</table>
Section 5.2 | The Regression Equation

Exercises

1) The association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending for a period of one year. The scatter plot below shows the relationship between these two variables.

(a) Describe the relationship between number of tourists and spending including direction, form, and strength.
(b) What are the explanatory and response variables? State the units for each variable.

2) The scatter plot below shows the relationship between the number of calories and amount of carbohydrates (in grams) for food menu items at Starbucks. Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbohydrates a menu item has based on its calorie content.

(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) for food menu items at Starbucks. Your description should include direction, form, and strength.
(b) What are the explanatory and response variables? State the units for each variable.

3) Data is collected on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The corresponding regression equation is $y = 51.32 + 0.725x$, where $x$ represents the distance in miles between two stops on the train, and $y$ represents the travel time in hours between the two stops.

(a) State the value of the slope of the regression equation and write a sentence to interpret this value.
(b) State the coordinates of the $y$-intercept of the regression equation and write a sentence to interpret this point.

4) Data is collected on the relationship between socioeconomic status, measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch), and the percentage of bike riders in the neighborhood wearing helmets (helmet). The regression line for this data is given by $helmet = 0.553 - 0.537 \times lunch$.

(a) State the value of the slope of the regression equation and write a sentence to interpret this value.
(b) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.
5) Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person has consumed. Sixteen student volunteers at Ohio State University each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The scatter plot is pictured below, and the regression equation is \( \hat{y} = -0.0127 + 0.0180x \).

(a) Describe the relationship between the number of cans of beer and BAC. Your description should include direction, form, and strength.

(b) State the value of the slope of the regression equation and write a sentence to interpret this value.

(c) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.

(d) Do you trust the regression equation to make fairly accurate predictions of BAC based on the number of cans of beer consumed? Explain why or why not.

6) Data is collected using a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years, to examine the relationship between the heights of married couples in that age range. The regression equation to predict the wife’s height in inches given her husband’s height in inches is \( \hat{y} = 43.5755 + 0.2863x \).

(a) State the value of the slope of the regression equation and write a sentence to interpret this value.

(b) State the coordinates of the y-intercept of the regression equation and write a sentence to interpret this point.

(c) Do you trust this regression equation to make fairly accurate predictions of a wife’s height based on her husband’s height? Explain why or why not.

7) A random sample of ten professional athletes produced the data below, where \( x \) is the number of endorsements the player has and \( y \) is the amount of money made (in millions of dollars).

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data.

(b) Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(c) State the value of the slope of the regression equation and write a sentence to interpret this value.

(d) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.
8) The table below shows the life expectancy in years for an individual born in the United States in certain years.

<table>
<thead>
<tr>
<th>Year of Birth</th>
<th>Life Expectancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1930</td>
<td>59.7</td>
</tr>
<tr>
<td>1940</td>
<td>62.9</td>
</tr>
<tr>
<td>1950</td>
<td>70.2</td>
</tr>
<tr>
<td>1965</td>
<td>69.7</td>
</tr>
<tr>
<td>1973</td>
<td>71.4</td>
</tr>
<tr>
<td>1982</td>
<td>74.5</td>
</tr>
<tr>
<td>1987</td>
<td>75</td>
</tr>
<tr>
<td>1992</td>
<td>75.7</td>
</tr>
<tr>
<td>2010</td>
<td>78.7</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data.

(b) Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(c) State the value of the slope of the regression equation and write a sentence to interpret this value.

(d) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.

9) The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition ten, for various pages is given in the table below.

<table>
<thead>
<tr>
<th>Page number</th>
<th>Maximum value ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>43</td>
<td>19</td>
</tr>
<tr>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>72</td>
<td>16</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data.

(b) Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(c) State the value of the slope of the regression equation and write a sentence to interpret this value.

(d) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.

10) The table below gives the gold medal times for every other Summer Olympics for the women’s 100-meter freestyle swimming event.

<table>
<thead>
<tr>
<th>Year</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1912</td>
<td>82.2</td>
</tr>
<tr>
<td>1924</td>
<td>72.4</td>
</tr>
<tr>
<td>1932</td>
<td>66.8</td>
</tr>
<tr>
<td>1952</td>
<td>66.8</td>
</tr>
<tr>
<td>1960</td>
<td>61.2</td>
</tr>
<tr>
<td>1968</td>
<td>60.0</td>
</tr>
<tr>
<td>1976</td>
<td>55.65</td>
</tr>
<tr>
<td>1984</td>
<td>55.92</td>
</tr>
<tr>
<td>1992</td>
<td>54.64</td>
</tr>
<tr>
<td>2000</td>
<td>53.8</td>
</tr>
<tr>
<td>2008</td>
<td>53.1</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data.

(b) Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(c) State the value of the slope of the regression equation and write a sentence to interpret this value.

(d) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.
11) We are interested in determining whether there is a linear relationship between the number of letters in a state name and the year the state entered the Union. Data for a selection of states is shown below.

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data.
(b) Use the applet to find the regression equation and include the line of best fit on the scatter plot.
(c) State the value of the slope of the regression equation and write a sentence to interpret this value.
(d) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.

<table>
<thead>
<tr>
<th>State</th>
<th># letters in name</th>
<th>Year entered the Union</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>7</td>
<td>1819</td>
</tr>
<tr>
<td>Colorado</td>
<td>8</td>
<td>1876</td>
</tr>
<tr>
<td>Hawaii</td>
<td>6</td>
<td>1959</td>
</tr>
<tr>
<td>Iowa</td>
<td>4</td>
<td>1846</td>
</tr>
<tr>
<td>Maryland</td>
<td>8</td>
<td>1788</td>
</tr>
<tr>
<td>Missouri</td>
<td>8</td>
<td>1821</td>
</tr>
<tr>
<td>New Jersey</td>
<td>9</td>
<td>1787</td>
</tr>
<tr>
<td>Ohio</td>
<td>4</td>
<td>1803</td>
</tr>
<tr>
<td>South Carolina</td>
<td>13</td>
<td>1788</td>
</tr>
<tr>
<td>Utah</td>
<td>4</td>
<td>1896</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>9</td>
<td>1848</td>
</tr>
</tbody>
</table>
5.3 | Prediction and Residuals

There are two types of predictions that we can make using a least squares regression equation, **interpolation** and **extrapolation**. The difference between these two types of predictions involves which value of the explanatory variable we use to make the prediction. If we use a value for the independent variable that is within the interval of the observed data, then we are using interpolation to make a prediction. If we use a value for the independent variable that is outside the interval of observed data, then we are using extrapolation to make a prediction. In the case of extrapolation, we need to be very careful to make sure the prediction makes sense. The regression equation may work reasonably well within the interval of the observed data but fail miserably outside that range. It will be your job to determine whether a predication made using extrapolation is valid.

Recall Example 5.2.1, which involved the linear relationship between students’ grades on their third exam and their grades on the final exam. Suppose you want to estimate, or predict, the final exam score of statistics students who received 73 on the third exam. The observed third exam scores (explanatory variable values) range from 65 to 75, and the observed final exam scores (response variable values) range between 126 and 198. Since \( x = 73 \) is within the interval of observed \( x \)-values, this is an example of interpolation. We substitute \( x = 73 \) into the regression equation and calculate \( \hat{y} \) as follows:

\[
\hat{y} = -173.51 + 4.83(73) = 179.08
\]

We predict that statistics student who earn a grade of 73 out of 80 on the third exam will earn a grade of about 179 out of 200 on the final exam, on average.

**NOTE**

The process of predicting values of \( y \) for \( x \)-values within the interval of observed \( x \)-values is called **interpolation**.

The process of predicting values of \( y \) for \( x \)-values outside the interval of observed \( x \)-values is called **extrapolation**.

Note that the interpretation of the vertical intercept may be extrapolation and may not make sense in context. For example, recall that in Section 5.1, we considered a scatter plot of height versus foot length. The observed foot lengths were all between 22 and 35 cm, and the vertical intercept of the regression equation was \((0, 38.3)\). This means that people with a shoe size of 0 are predicted to be 38.3 inches tall, on average. This does not make sense, since a shoe size of 0 is unrealistic. In addition, a shoe size of 0 is outside the observed interval of the data, which makes the interpretation an example of extrapolation.
When we use a regression equation to make a prediction, it is extremely rare for the regression equation prediction to match up perfectly with an observed value of the response variable. The regression equation is a model that can be used to make a prediction, but we need to keep in mind that there will be differences between what the model predicts and the true values of the response variable.

For example, in Example 5.2.1 one original observed data point is (71, 185). This point was the observation of a student who earned 71 out of 80 on the third exam and earned 185 out of 200 on the final exam. As shown below, the regression model would predict a different final exam score for this student.

\[ \hat{y} = -173.51 + 4.83 \times 71 = 169.42 \]

The regression equation predicts that a student who earns a score of 71 on the third exam would, on average, earn a score of about 169 on the final exam. In this case, the regression equation underestimates the observed final exam score by about 16 points, since \( 185 - 169.42 = 15.58 \).

This difference between the observed final exam score and the predicted final exam score from the regression equation is called the residual.

**NOTE**

The residual is the difference between the observed output value and the predicted output value based on the regression equation:

\[ \text{residual} = (\text{observed true value of output}) - (\text{predicted value of output from regression equation}) \]

A positive residual means the observed output value is above the regression line, and the regression equation produced an underestimate of the true value.

A negative residual means the observed output value is below the regression line, and the regression equation produced an overestimate of the true value.
Example 5.3.1

Recall Example 5.2.1 regarding predicting students’ final exam scores from their scores on the third exam. The regression equation is \( \hat{y} = -173.51 + 4.83x \), where \( x \) is the score (out of 80 points) on the third exam and \( y \) is the final exam score (out of 200 points).

(a) What would you predict the final exam score to be for a student who scored a 65 on the third exam?
(b) What would you predict the final exam score to be for a student who scored a 78 on the third exam?
(c) Given the fact that one student in the class was observed to have earned a 65 on the third exam and then earned a 175 on the final exam, find the residual for the student and interpret the residual value in context.

Solution 5.3.1

(a) Replacing \( x \) in the regression equation with 65 we calculate the predicted final exam score as follows:

\[
\hat{y} = -173.51 + 4.83(65) = 140.44
\]

A student who earned a 65 on the third exam is predicted to earn a score of 140.44 on the final exam.

(b) The \( x \) values in the data set are between 65 and 75 out of a total of 80 points possible on the exam. A score of 78 is outside the interval of observed \( x \) values in the data set. Using the regression equation to make this prediction is called extrapolation. You cannot reliably predict the final exam score for this student. Even though it is possible to enter 78 into the equation for \( x \) and calculate a corresponding \( \hat{y} \) value, the predicted \( y \) value that you calculate may not be reliable.

To understand how unreliable the prediction can be when you use a value of the explanatory variable that is outside the interval of observed \( x \) values, make the substitution \( x = 78 \) into the equation.

\[
\hat{y} = -173.51 + 4.83(78) = 203.23
\]

The final exam score is predicted to be 203.23. But, the largest the final exam score can be is 200, so this answer does not make sense in context. Remember, when using extrapolation, we can't necessarily trust the answers we get from the equation!

(c) The observed final exam score for the student who earned 65 on the third exam was 175. The predicted final exam score based on the regression equation is 140.44. The residual is the difference between these two values, in the order residual = (observed value) – (predicted value). In this case, the residual is

\[
175 - 140.44 = 34.56
\]

This means that the observed final exam score was almost 35 points higher than the final exam score predicted by the regression equation. In other words, the regression model underestimated the final exam score by about 35 points.
5.3.1

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test (out of 100 points). The equation for the line of best fit is

\[ \hat{y} = 72.5 + 2.8x, \]

where \( x \) is the number of hours per week spent practicing and \( y \) is the score on the math test.

(a) What is the predicted score on the math test for a student who practices a musical instrument for five hours a week?

(b) If we observed that a student who practiced five hours in a week earned a score of 75 on the math test, then find the value of the residual and interpret the residual.
Example 5.3.2

The scatter plot below shows data on head length (in mm) versus total length (in cm) for a sample of 104 brushtail possums. The equation of best-fit for this data set is \( \hat{y} = 41 + 0.59x \).

![Scatter plot showing head length vs. total length for 104 brushtail possums.](https://www.flickr.com/photos/wollombi/58499575)

The common brushtail possum of Australia. Photo by wollombi on Flickr: [www.flickr.com/photos/wollombi/58499575](https://www.flickr.com/photos/wollombi/58499575)

A point representing a possum with head length 94.1 mm and total length 89 cm is highlighted on the scatter plot above.

Since total length is shown along the horizontal axis in the scatter plot, we know that total length is the input (explanatory) variable and is denoted by \( x \) in the regression equation. Since head length is shown along the vertical axis, we know that head length is the output (response) variable and is denoted by \( y \) in the regression equation.

To predict the head length of a possum that has a total length of 80 cm we can substitute 80 for \( x \) in the regression equation and calculate \( \hat{y} = 41 + 0.59 \times 80 = 88.2 \). In other words, a possum with a total length of 80 cm is predicted to have a head length of 88.2 mm, on average.

Recall that the “hat” on \( y \) is used to signify that the linear equation gives a predicted head length. This estimate can be viewed as an average in that the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

If we were to now observe a possum that had a total length of 80 cm and a head length of 90 mm, then we could also calculate the residual:

\[
\text{residual} = \text{(observed value) } - \text{(predicted value)} \\
\text{residual} = (90 \text{ mm observed length}) - (88.2 \text{ mm predicted length}) \\
\text{residual} = 1.8 \text{ mm}
\]

This tells us that the observed head length of the possum was 1.8 mm larger than the value predicted by the regression equation. The regression equation underestimated the true value of this possum's head length.
Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks, the climate debate rages on.

Stephen Colbert
April 6, 2010
Source

Linear models can be used to approximate the relationship between two quantitative variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

Example 5.3.3

The amount gifted to the family, in thousands of dollars, and the family income, in thousands of dollars, is collected from a sample of 50 students at Elmhurst College. The points are shown in the scatter plot at right, together with the least squares regression line, $\hat{y} = 24.3 - 0.0431x$.

(a) Identify (describe in words) the explanatory and response variables.
(b) State the value of the slope and interpret the slope in context.
(c) State the vertical intercept and interpret this point in context.
(d) Use the regression equation to predict the amount of gift aid students receive if their family income is $1 million. Is this an example of interpolation or extrapolation? Explain.

Solution 5.3.1

(a) The explanatory variable is family income in thousands of dollars and is seen on the horizontal axis in the scatter plot. The response variable is the amount of gift aid from the university in thousands of dollars and is seen on the vertical axis in the scatter plot.
(b) The value of the slope is $-0.0431$. This means that for each $1000$ increase in family income, the amount of gift aid from the university is predicted to decrease by about $43.10 (0.0431\times1000)$, on average.
(c) The vertical intercept is $(0, 24.3)$. This means that for families with $0$ income, the amount of gift aid from Elmhurst College is predicted to be $24,300$ on average. Since the scatter plot shows points that appear to have family incomes close to $0$, this may be a reasonable estimate even though it is an extrapolation.

(continued on next page)
Example 5.3.3 continued

Solution 5.3.3 continued

(d) Since the units of family income are thousands of dollars, we write $1 million as $1000 thousand. We then substitute \( x = 1000 \) into the regression equation and calculate
\[
\hat{y} = 24.3 - 0.0431 \times 1000 = -18.8.
\]
The model predicts that students with family income of $1 million are predicted to receive an average of \(-18.8\) thousand dollars in gift aid. The scatter plot shows that $1 million is outside the interval of observed values of the explanatory variable (family income values appear to range from near $0 to almost $300,000). Therefore, this prediction is an example of extrapolation. In addition, the prediction doesn’t make sense, since a negative amount of gift aid is not realistic. The linear regression model can be expected to provide reasonable predictions for family income values like those observed in the data. Since the relationship between family income and gift aid has not been explored for family income values above $300,000, we cannot expect the regression equation to give a valid prediction of gift aid for a family income of $1 million.
Section 5.3 | Prediction and Residuals

Exercises

1) Data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles is collected, and the travel time between two stops on the train (in minutes) is modeled by the equation \( \hat{y} = 51.320.725x \), where \( x \) is the number of miles between the two stops. Source: Starbucks.com, collected on March 10, 2011, http://www.starbucks.com/menu/nutrition

(a) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(b) In reality, it takes the Coast Starlight about 168 minutes to travel from Santa Barbara to Los Angeles. Calculate the residual, and explain the meaning of this value.

2) Data on shoulder girth and height of a group of individuals is collected. The regression equation for these variables is \( \hat{y} = 105.36 + 0.608x \), where \( x \) is the shoulder girth in cm, and \( y \) is the height of the individual in cm.

(a) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the regression model.

(b) The student from part (a) is actually 160 cm tall. Calculate the residual, and explain the meaning of this value.

(c) A one-year-old has a shoulder girth of 56 cm. Would it be appropriate to use the linear model to predict the height of this child? Explain.

3) The scatter plot below shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (\( \text{lunch} \)), and the percentage of bike riders in the neighborhood wearing helmets (\( \text{helmet} \)). The regression line for this data is given by \( \text{helmet} = 0.553 - 0.537 \times \text{lunch} \). What would the value of the residual be for a neighborhood where 50% of the children receive reduced-free lunches, and 40% of the bike riders wear helmets in that neighborhood? Interpret the meaning of this residual in context.
4) Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person has consumed. Sixteen student volunteers at Ohio State University each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The scatter plot is pictured below, and the regression equation is \( \hat{y} = -0.0127 + 0.0180x \).

(a) One of the volunteers in the Ohio State University study drank 5 cans of beer and had a blood alcohol content of .10 grams of alcohol per deciliter. Calculate the residual and explain the meaning of the residual value.

(b) One of the volunteers in the Ohio State University study drank 5 beers and had a blood alcohol content of 0.05 grams of alcohol per deciliter. Calculate the residual and explain the meaning of the residual value.

5) The scatter plot below shows the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals. The regression equation for this data is \( \hat{y} = -105.0113 + 1.0176x \).

(a) One of the people in the sample was 164 cm tall and weighed 82 kg. Calculate the residual and explain the meaning of the residual value.

(b) One of the people in the sample was 181 cm tall and weighed 65 kg. Calculate the residual and explain the meaning of the residual value.

6) The scatter plot displays the relationship between husbands’ and wives’ ages in a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years. The regression equation is \( \hat{y} = 1.5740 + 0.9112x \).

(a) One of the couples in the sample had a husband who was 40 years old and a wife who was 30 years old. Calculate the residual and explain the meaning of the residual value.

(b) One of the couples in the sample had a husband who was 20 years old and a wife who was 22 years old. Calculate the residual and explain the meaning of the residual value.
7) The scatter plot below summarizes husbands’ and wives’ heights in a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years. The regression equation is $\hat{y} = 43.5755 + 0.2863x$.

(a) One of the couples in the sample had a husband who was 70 inches tall and a wife who was 57 inches tall. Calculate the residual and explain the meaning of the residual value.

(b) One of the couples in the sample had a husband who was 70 inches tall and a wife who was 65 inches tall. Calculate the residual and explain the meaning of the residual value.

8) A random sample of ten professional athletes produced the data below, where $x$ is the number of endorsements the player has and $y$ is the amount of money made (in millions of dollars).

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) According to the regression model, how much money would a player make (in millions of dollars) if the player had 4 endorsements?

(c) Calculate the residual when $x = 4$. Write a sentence to interpret the meaning of the residual value.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

9) An electronics retailer used regression to find a model to predict sales in the first quarter of the new year (January through March). The model is good for 90 days, where $x$ is the day. The model can be written as follows: $\hat{y} = 101.32 + 2.48x$, where $y$ is in thousands of dollars.

(a) What does the model predict sales to be on day 60? When collecting the data, the retailer’s observed data point for $x = 60$ had a residual of 10. Identify the sales figure that was observed by the retailer for day 60.

(b) What does the model predict sales to be on day 90? When collecting the data, the retailer’s observed data point for $x = 90$ had a residual of $-13$. Identify the sales figure that was observed by the retailer for day 90.

10) A landscaping company is hired to mow the grass for several large properties. The total area of all the properties combined is 1,345 acres. The number of acres left to mow is estimated by the equation $\hat{y} = 1350 - 1.2x$, where $x$ is the number of hours worked.

(a) According to the model, how many acres will be left to mow after 20 hours of work? 

(b) According to the model, how many acres will be left to mow after 100 hours of work?

(c) According to the model, how many hours will it take to mow all of the lawns?
11) The table below contains real data for the first two decades of flu cases reporting.

<table>
<thead>
<tr>
<th>Year</th>
<th># flu cases diagnosed</th>
<th># flu deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1981</td>
<td>91</td>
<td>29</td>
</tr>
<tr>
<td>1981</td>
<td>319</td>
<td>121</td>
</tr>
<tr>
<td>1982</td>
<td>1,170</td>
<td>453</td>
</tr>
<tr>
<td>1983</td>
<td>3,076</td>
<td>1,482</td>
</tr>
<tr>
<td>1984</td>
<td>6,240</td>
<td>3,466</td>
</tr>
<tr>
<td>1985</td>
<td>11,776</td>
<td>6,878</td>
</tr>
<tr>
<td>1986</td>
<td>19,032</td>
<td>11,987</td>
</tr>
<tr>
<td>1987</td>
<td>28,564</td>
<td>16,162</td>
</tr>
<tr>
<td>1988</td>
<td>35,447</td>
<td>20,868</td>
</tr>
<tr>
<td>1989</td>
<td>42,674</td>
<td>27,591</td>
</tr>
<tr>
<td>1990</td>
<td>48,634</td>
<td>31,335</td>
</tr>
<tr>
<td>1991</td>
<td>59,660</td>
<td>36,560</td>
</tr>
<tr>
<td>1992</td>
<td>78,530</td>
<td>41,055</td>
</tr>
<tr>
<td>1993</td>
<td>78,834</td>
<td>44,730</td>
</tr>
<tr>
<td>1994</td>
<td>71,874</td>
<td>49,095</td>
</tr>
<tr>
<td>1995</td>
<td>68,505</td>
<td>49,456</td>
</tr>
<tr>
<td>1996</td>
<td>59,347</td>
<td>38,510</td>
</tr>
<tr>
<td>1997</td>
<td>47,149</td>
<td>20,736</td>
</tr>
<tr>
<td>1998</td>
<td>38,393</td>
<td>19,005</td>
</tr>
<tr>
<td>1999</td>
<td>25,174</td>
<td>18,454</td>
</tr>
<tr>
<td>2000</td>
<td>25,522</td>
<td>17,347</td>
</tr>
<tr>
<td>2001</td>
<td>25,643</td>
<td>17,402</td>
</tr>
<tr>
<td>2002</td>
<td>26,464</td>
<td>16,371</td>
</tr>
<tr>
<td>Total</td>
<td>802,118</td>
<td>489,093</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Do not include pre-1981 data.

(b) Use the applet to find the regression equation. Round values in the equation to the nearest whole number.

(c) Include the line of best fit on the scatter plot. Does the line seem to fit the data? Explain.

(d) Use the model to find the predicted number of flu cases diagnosed in 1970. Why doesn't this answer make sense?

12) The table below shows the life expectancy for an individual born in the United States in certain years.

<table>
<thead>
<tr>
<th>Year of Birth</th>
<th>Life Expectancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1930</td>
<td>59.7</td>
</tr>
<tr>
<td>1940</td>
<td>62.9</td>
</tr>
<tr>
<td>1950</td>
<td>70.2</td>
</tr>
<tr>
<td>1965</td>
<td>69.7</td>
</tr>
<tr>
<td>1973</td>
<td>71.4</td>
</tr>
<tr>
<td>1982</td>
<td>74.5</td>
</tr>
<tr>
<td>1987</td>
<td>75</td>
</tr>
<tr>
<td>1992</td>
<td>75.7</td>
</tr>
<tr>
<td>2010</td>
<td>78.7</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.

(c) Why aren't the answers to part (b) the same as the values in the table that correspond to those years?

(d) Find the residuals for the values found in part (b), and then interpret the meanings of the residual values in context.
13) The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition ten, for various pages is given in the table below.

<table>
<thead>
<tr>
<th>Page number</th>
<th>Maximum value ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>43</td>
<td>19</td>
</tr>
<tr>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>72</td>
<td>16</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) Find the estimated maximum values for the restaurants on page 10 and on page 70.

(c) Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?

(d) Is the least squares line valid for page 200? Why or why not?

14) The table below gives the gold medal times for every other Summer Olympics for the women’s 100-meter freestyle swimming event.

<table>
<thead>
<tr>
<th>Year</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1912</td>
<td>82.2</td>
</tr>
<tr>
<td>1924</td>
<td>72.4</td>
</tr>
<tr>
<td>1932</td>
<td>66.8</td>
</tr>
<tr>
<td>1952</td>
<td>66.8</td>
</tr>
<tr>
<td>1960</td>
<td>61.2</td>
</tr>
<tr>
<td>1968</td>
<td>60.0</td>
</tr>
<tr>
<td>1976</td>
<td>55.65</td>
</tr>
<tr>
<td>1984</td>
<td>55.92</td>
</tr>
<tr>
<td>1992</td>
<td>54.64</td>
</tr>
<tr>
<td>2000</td>
<td>53.8</td>
</tr>
<tr>
<td>2008</td>
<td>53.1</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) Find the estimated gold medal time for 1932. Find the estimated time for 1984.

(c) Why are the answers from part (b) different from the values in the table?

(d) Does it appear that a line is the best model to fit the data? Why or why not?

(e) Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?
15) We are interested in determining whether there is a linear relationship between the number of letters in a state name and the year the state entered the Union. Data for a selection of states is shown below.

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.
(b) Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
(c) Does it appear that a line is the best model to fit the data? Why or why not?
(d) Use the least-squares line to estimate the number of letters a new state that enters the Union this year would have. Should the least-squares line be used to predict this? Why or why not?

<table>
<thead>
<tr>
<th>State</th>
<th># letters in name</th>
<th>Year entered the Union</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>7</td>
<td>1819</td>
</tr>
<tr>
<td>Colorado</td>
<td>8</td>
<td>1876</td>
</tr>
<tr>
<td>Hawaii</td>
<td>6</td>
<td>1959</td>
</tr>
<tr>
<td>Iowa</td>
<td>4</td>
<td>1846</td>
</tr>
<tr>
<td>Maryland</td>
<td>8</td>
<td>1788</td>
</tr>
<tr>
<td>Missouri</td>
<td>8</td>
<td>1821</td>
</tr>
<tr>
<td>New Jersey</td>
<td>9</td>
<td>1787</td>
</tr>
<tr>
<td>Ohio</td>
<td>4</td>
<td>1803</td>
</tr>
<tr>
<td>South Carolina</td>
<td>13</td>
<td>1788</td>
</tr>
<tr>
<td>Utah</td>
<td>4</td>
<td>1896</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>9</td>
<td>1848</td>
</tr>
</tbody>
</table>

16) A researcher is investigating the relationship between white male population size and homicide rate. He uses demographic data from Detroit, MI to compare homicide rates to the number of the population that are white males. The data are shown in the table below.

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.
(b) State the value of the slope of the regression equation and write a sentence to interpret this value.
(c) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.
(d) Which observed data value has the largest residual? Explain the meaning of the residual in context.

<table>
<thead>
<tr>
<th>Population Size</th>
<th>Homicide rate per 100,000 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>558,724</td>
<td>8.6</td>
</tr>
<tr>
<td>538,584</td>
<td>8.9</td>
</tr>
<tr>
<td>519,171</td>
<td>8.52</td>
</tr>
<tr>
<td>500,457</td>
<td>8.89</td>
</tr>
<tr>
<td>482,418</td>
<td>13.07</td>
</tr>
<tr>
<td>465,029</td>
<td>14.57</td>
</tr>
<tr>
<td>448,267</td>
<td>21.36</td>
</tr>
<tr>
<td>432,109</td>
<td>28.03</td>
</tr>
<tr>
<td>416,533</td>
<td>31.49</td>
</tr>
<tr>
<td>401,518</td>
<td>37.39</td>
</tr>
<tr>
<td>387,046</td>
<td>46.26</td>
</tr>
<tr>
<td>373,095</td>
<td>47.24</td>
</tr>
<tr>
<td>359,647</td>
<td>52.33</td>
</tr>
</tbody>
</table>
17) The height in feet (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). The data are shown in the table below.

<table>
<thead>
<tr>
<th>Height (in feet)</th>
<th>Stories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,050</td>
<td>57</td>
</tr>
<tr>
<td>428</td>
<td>28</td>
</tr>
<tr>
<td>362</td>
<td>26</td>
</tr>
<tr>
<td>529</td>
<td>40</td>
</tr>
<tr>
<td>790</td>
<td>60</td>
</tr>
<tr>
<td>401</td>
<td>22</td>
</tr>
<tr>
<td>380</td>
<td>38</td>
</tr>
<tr>
<td>1,454</td>
<td>110</td>
</tr>
<tr>
<td>1,127</td>
<td>100</td>
</tr>
<tr>
<td>700</td>
<td>46</td>
</tr>
</tbody>
</table>

(a) Using “Stories” as the independent variable and “Height” as the dependent variable, use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) Find the predicted heights of buildings with 32 stories and with 94 stories.

(c) Does it appear that there is a linear relationship between the number of stories in tall buildings and the height of the buildings? Explain.

(d) What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of the height of a building with six stories? Explain.

(e) Based on the regression line equation, adding an extra story is predicted to add about how many feet to a building?

18) The table below shows data on average per capita coffee consumption and heart disease rate in a random sample of 10 countries.

<table>
<thead>
<tr>
<th>Yearly coffee consumption in liters</th>
<th>Death from heart diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>221</td>
</tr>
<tr>
<td>3.9</td>
<td>167</td>
</tr>
<tr>
<td>2.9</td>
<td>131</td>
</tr>
<tr>
<td>2.4</td>
<td>191</td>
</tr>
<tr>
<td>2.9</td>
<td>220</td>
</tr>
<tr>
<td>0.8</td>
<td>297</td>
</tr>
<tr>
<td>9.1</td>
<td>71</td>
</tr>
<tr>
<td>2.7</td>
<td>172</td>
</tr>
<tr>
<td>0.8</td>
<td>211</td>
</tr>
<tr>
<td>0.7</td>
<td>300</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) State the value of the slope of the regression equation and write a sentence to interpret this value.

(c) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.

(d) How well does the regression line fit the data? Explain.

(e) Which observed data value has the largest residual? Explain what the residual means in context.

19) The table below shows one student athlete’s time (in minutes) to swim 2000 yards and the student’s heart rate (beats per minute) after swimming on a random sample of 10 days.

<table>
<thead>
<tr>
<th>Swim Time</th>
<th>Heart Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.12</td>
<td>144</td>
</tr>
<tr>
<td>35.72</td>
<td>152</td>
</tr>
<tr>
<td>34.72</td>
<td>124</td>
</tr>
<tr>
<td>34.05</td>
<td>140</td>
</tr>
<tr>
<td>34.13</td>
<td>152</td>
</tr>
<tr>
<td>35.73</td>
<td>146</td>
</tr>
<tr>
<td>36.17</td>
<td>128</td>
</tr>
<tr>
<td>35.57</td>
<td>136</td>
</tr>
<tr>
<td>35.37</td>
<td>144</td>
</tr>
<tr>
<td>35.57</td>
<td>148</td>
</tr>
</tbody>
</table>

(a) Use the Two Quantitative Variables applet to create a scatter plot of the data. Use the applet to find the regression equation and include the line of best fit on the scatter plot.

(b) State the value of the slope of the regression equation and write a sentence to interpret this value.

(c) State the coordinates of the vertical intercept of the regression equation and write a sentence to interpret this point.

(d) How well does the regression line fit the data? Explain.

(e) Which observed data value has the largest residual? Explain what the residual means in context.
5.4 | Testing the Significance of the Correlation Coefficient

The Correlation Coefficient $r$

Besides looking at the scatter plot and seeing that a line seems like a reasonable model for the data, how can you tell if the line is a good predictor? The correlation coefficient is another indicator (besides the scatterplot) of the strength of the relationship between $x$ and $y$.

The correlation coefficient, $r$ (developed by Karl Pearson in the early 1900s) provides a measure of strength and direction of the linear association between the independent variable $x$ and the dependent variable $y$.

If you suspect a linear relationship between $x$ and $y$, then the correlation coefficient, $r$, can measure how strong the linear relationship is.

Facts about $r$:

- The value of $r$ is always between $-1$ and 1. $-1 \leq r \leq 1$
- The size of the correlation coefficient $r$ indicates the strength of the linear relationship between $x$ and $y$. Values of $r$ close to -1 or close to +1 indicate a stronger linear relationship between $x$ and $y$.
  - If $r = 0$, there is no linear relationship between $x$ and $y$ (no linear correlation).
  - If $r = 1$, there is perfect positive correlation. All data points fall perfectly on a straight, increasing line.
  - If $r = -1$, there is perfect negative correlation. All data points fall perfectly on a straight, decreasing line.
- A positive value of $r$ means that when $x$ increases, $y$ tends to increase and when $x$ decreases, $y$ tends to decrease (positive correlation).
- A negative value of $r$ means that when $x$ increases, $y$ tends to decrease and when $x$ decreases, $y$ tends to increase (negative correlation).
- The sign of $r$ is the same as the sign of the slope of the best-fit line.

**NOTE**

Strong correlation does not suggest that $x$ causes $y$ or that $y$ causes $x$. Correlation does not imply causation!
Correlation Guessing Game

The following three scatterplots were created by the Correlation Guessing Game Applet.

Guess the correlation of each scatter plot using the information above! Answers are on the next page. Don't cheat!
*Answers to the Correlation Guessing Game: Your guesses should be near these values.
   a) \( r = -0.8 \)  
   b) \( r = 0.7 \)  
   c) \( r = -0.7 \)

Example 5.4.1

(a) A scatter plot with positive correlation 
(b) A scatter plot with negative correlation. 
(c) A scatter plot showing data with zero correlation.

The correlation coefficient, \( r \), tells us about the strength and direction of the linear relationship between the explanatory and response variables in the sample data. However, the reliability of the linear model also depends on how many observed data points are in the sample.

We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute \( r \), the correlation coefficient for the sample. If we had data for the entire population, then we could find the population correlation coefficient. But because we only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, \( r \), is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient value is the Greek letter \( p \), pronounced "rho."

\[
\begin{align*}
   p & = \text{population correlation coefficient} \quad \text{(unknown)} \\
   r & = \text{sample correlation coefficient value} \quad \text{(known; calculated from sample data)}
\end{align*}
\]
The hypothesis test lets us decide whether the value of the population correlation coefficient, \( p \), is "close to zero" (meaning there is no correlation between \( x \) and \( y \) in the population) or if \( p \) is "significantly different from zero" (meaning there is correlation between \( x \) and \( y \) in the population).

*If the hypothesis test indicates that the population correlation coefficient is significantly different from zero, then we conclude that there is statistically significant evidence of correlation between \( x \) and \( y \) in the population.*

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between \( x \) and \( y \) because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a statistically significant relationship between \( x \) and \( y \). We can use the regression line to model the linear relationship between \( x \) and \( y \) in the population.

*If the hypothesis test indicates that the population correlation coefficient is NOT significantly different from zero (i.e. it is close to zero), then we conclude that it is plausible that there is no correlation between \( x \) and \( y \) in the population.*

- Conclusion: There is insufficient evidence to conclude that there is a significant linear relationship between \( x \) and \( y \) because the correlation coefficient is not significantly different from zero.
- What the conclusion means: There is not a statistically significant relationship between \( x \) and \( y \). We CANNOT use the regression line to model a linear relationship between \( x \) and \( y \) in the population.

**Performing a Hypothesis Test for Evidence of Correlation**

To perform a hypothesis test for significant evidence of correlation, the hypotheses are always as follows:

\[ H_0: \rho = 0 \]  
There is no linear correlation between the explanatory and response variables in the population. The population correlation coefficient is equal to 0.

\[ H_1: \rho \neq 0 \]  
There is a linear correlation between the explanatory and response variables in the population. The population correlation coefficient is different from 0.

After writing the hypotheses, you will then use the [Least Squares Regression Applet](#) to find the simulation-based \( p \)-value (see instructions below). To make the conclusion:

- If the \( p \)-value is less than the significance level, then we reject the null hypothesis. This means the alternative hypothesis is plausible. There is statistically significant evidence of a linear correlation between the explanatory and response variables in the population because the correlation coefficient is significantly different from 0.

- If the \( p \)-value is more than the significance level, then we fail to reject the null hypothesis. This means the null hypothesis is plausible. There is insufficient evidence to conclude that there is a significant linear correlation between the explanatory and response variables in the population because the correlation coefficient is not significantly different from 0.
Using Simulation to Test for Significant Correlation

1. Enter the observed data into the Rossman Chance Least Squares Regression applet.

2. Select Show Shuffle Options, and select the Correlation Coefficient button.

3. Press the Shuffle Y-values option. The y-values from the sample data set will be randomly shuffled and re-assigned to the x-values from the sample data set.

   The resulting sample correlation coefficient will be displayed and plotted in the graph of sample correlation coefficients. Notice that this is shuffling the y-values randomly because the x-values and y-values would be randomly assigned to one another if the data was truly a data set with no correlation, which is what we are assuming from the null hypothesis. See this video for more on this!

4. Repeatedly press the Shuffle Y-values button, or increase the Number of Shuffles in order to produce a large number of shuffled correlation coefficients all at once. Generally, we jump to doing 1000 shuffles in the “Number of shuffles” box.

5. Find the simulated p-value using Count Shuffles. Since the alternative hypothesis is two-sided, we will always use the "Beyond" option for the count. Put the sample statistic, which is the sample correlation coefficient r, into the Count box.
Example 5.4.2

In real-world situations, it is extremely rare that we would have an entire population data set, and so it would be extremely rare that we would ever be able to view a scatterplot for the entire population and calculate the population correlation coefficient. However, for the sake of learning, this example gives the scatterplot for entire population data set, and the value of the population correlation coefficient $\rho$ (pronounced "rho") is $\rho = -0.9647$.

Suppose that we have three different people who are trying to understand whether there is correlation between the variables $x$ and $y$. Remember, these three people wouldn't actually have the population data or the population correlation coefficient like we see above. Instead, they would only be able to take a random sample from the population, identify the correlation coefficient of their own sample data, and then use that sample correlation coefficient to make a strength of evidence conclusion about the correlation in the population. In the following pages, we will explore the three samples that the three people gathered, and the conclusions they would make about the population.

(continued on the following pages)
Person #1: Suppose that Person #1 collects a random sample of 34 points from the population. A scatterplot of the sample data, is shown below, and the sample correlation coefficient is \( r_1 = -0.986 \).

Person #1 would then conduct a hypothesis test to try to determine whether there is strong evidence of correlation between x and y in the population.

Person #1 would first create null and alternative hypotheses.

\[ H_0: \rho = 0 \quad \text{The null hypothesis is that there is no correlation between x and y in the population; the population correlation coefficient is 0.} \]

\[ H_A: \rho \neq 0 \quad \text{The alternative hypothesis is that there is correlation between x and y in the population; the population correlation coefficient is not 0.} \]

Person #1 would assume that the null hypothesis is true, and then determine the probability of getting a sample correlation coefficient at least as extreme as the sample correlation coefficient \( r_1 = -0.986 \). In other words, Person #1 would find a p-value for the sample observed. In order to do that, Person #1 can use the applet to create a distribution of simulated sample correlation coefficients:

1. The sample data is copied into the Rossman Chance Applet: [Least Squares Regression applet](https://www.rossmanchance.com/applets/LSRegressionApplet/).

2. The Show Shuffle Options box is checked, and the Correlation Coefficient button is selected.

3. The Number of Shuffles is set to 1000 in order to produce a large number of shuffled correlation coefficients all at once. We now have a simulated distribution of sample correlation coefficients. This distribution shows us the sample correlation coefficients that we would expect to see if there was no correlation in the population, as we assumed in the null hypothesis. The distribution is centered around 0, with some of the sample correlation coefficients slightly more than 0, and some slightly less than 0. But, we will find that there are not many simulated sample correlation coefficients that are much more than 0 or much less than 0 since that would be an unusual thing to happen if there were truly no correlation in the population.

4. Person #1 would find the simulated p-value for the observed sample data using Count Shuffles. Person #1 would select "Beyond" and enter the sample correlation coefficient value, \( r_1 = -0.986 \) in the Count box, and then press Count in order to find the p-value.

Person #1 had sample data that produced a simulated p-value of 0. Person #1 would be able to make the following conclusions:

- If there is no correlation between x and y in the population, then the probability of getting a sample data set with a correlation coefficient at least as extreme as \( r_1 = -0.986 \) is 0.

- Person #1 has a sample data set that gives strong evidence that there is correlation between x and y in the population.
Person #2: Now we will consider a second person who is also pulling a random sample of 34 points from the same population. The scatterplot of the sample data randomly collected by Person #2 is shown below, and the sample correlation coefficient for this data is $r^2 = -0.980$.

Person #2 conducts a hypothesis test to determine whether there is strong evidence of correlation between $x$ and $y$ in the population.

Person #3: Finally we will consider a third person who is also pulling a random sample of 34 points from the same population. The scatterplot of the sample data randomly collected by Person #3 is shown below, and the sample correlation coefficient for this data is $r^3 = -0.984$.

Person #3 conducts a hypothesis test to determine whether there is strong evidence of correlation between $x$ and $y$ in the population.

Person #2 had sample data that produced a simulated $p$-value of 0. Person #2 would be able to make the following conclusions:

- If there is no correlation in the population between $x$ and $y$, then the probability of getting a sample data set with a correlation coefficient at least as extreme as $r^2 = -0.980$ is 0.
- Person #2 has a sample data set that gives strong evidence that there is correlation between $x$ and $y$ in the population.

Person #3 had sample data that produced a simulated $p$-value of 0. Person #3 would be able to make the following conclusions:

- If there is no correlation in the population between $x$ and $y$, then the probability of getting a sample data set with a correlation coefficient at least as extreme as $r^3 = -0.984$ is 0.
- Person #3 has a sample data set that gives strong evidence that there is correlation between $x$ and $y$ in the population.

Conclusions we draw from exploring the three people’s random samples:

- Each of three people collected different points in their random samples, and therefore each person had a different sample correlation coefficient.
- The sample correlation coefficients, $r_1 = -0.986$, $r_2 = -0.980$, and $r_3 = -0.984$ were each "similar" to the true population parameter, $\rho = -0.9647$.
- Even though each of the three people had different sample correlation coefficients, each person made the same conclusion: that there is strong evidence of correlation in the population.
Example 5.4.3

Recall the Foot Length vs Height example: the line of best fit is \( \hat{y} = 38.30 + 1.03x \) with \( r = 0.711 \) and there are \( n=20 \) data points. Can the regression line be used for prediction? Given a foot length, can we predict a person's height using this equation? In other words, do the sample data provide evidence of correlation between height and foot length in the population?

**Conducting the hypothesis test:**

\[
H_0: \rho = 0 \quad \text{No correlation between foot length and height in the population.}
\]
\[
H_1: \rho \neq 0 \quad \text{There is correlation between foot length and height in the population.}
\]
\[
\alpha = 0.05
\]

- Enter the data in the applet.
- Select "Show Shuffle Options"
- Select "Correlation Coefficient"
- Type in a large "Number of Shuffles" (100 shuffles shown here)
- Click “Shuffle Y-values”

To find the p-value from this simulation, we will type the statistic, \( r=0.711 \) into the "Count Shuffles" box, choose the correct direction, which is “Beyond” and click "Count"!

The p-value is 0.

This is very strong evidence against the null!

Because the sample correlation coefficient, \( r \), is significantly different than 0, we conclude that we have strong evidence that there is a linear correlation between a person's foot length and a person's height.

There is strong evidence that there is a linear trend in the population. Therefore, the regression line equation can be used to predict height!
Using Applet to Find Theory-based p-value and t-score

We will continue using the foot-length and height data in order to demonstrate how to find the p-value and t-score (standardized statistic) using a theory-based approach instead of a simulation-based approach.

In order to find the t-score with a theory-based approach, select the Regression Table option in the applet.

The t-score associated with this sample data is given in the table under the t-stat column, in the 2nd row.

In this example, the t-score is 4.29, which is extremely strong evidence of correlation between foot length and height in the population. This matches our conclusion when we used the p-value from a simulation-based approach.

In order to use the applet to find the p-value using a theory based approach, first follow the instructions above in order to find the t-score from the theory-based approach. Then use that t-score in the applet, as described below, in order to find the p-value.

In order to find the p-value with a theory-based approach, select the t-statistic option in the applet.

Also select the Overlay t-distribution option in the applet.

Lastly, copy the t-score into the Count shuffles area of the applet and click Count.

In this example, the p-value from the theory-based approach is 0.0004 which is extremely strong evidence of correlation between foot length and height in the population. This matches our conclusion when we used the p-value from a simulation-based approach.

Note that the “Overlay t-distribution” option overlays a t-distribution on top of the distribution that was created using a simulation-based approach. If the appropriate theory-based conditions are met (see next page), then the t-distribution will be roughly bell shaped, centered at 0, and will create a t-score and p-value that matches the conclusion we found using a simulation-based approach.
Section 5.4 | Testing Correlation Significance
Written Homework

1) For each of the six scatterplots, estimate the value of the correlation coefficient.

![Scatterplots](image)

2) For each of the six scatterplots, estimate the value of the correlation coefficient.

![Scatterplots](image)

3) The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

(a) Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.

(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?

![Scatterplots](image)
4) Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.

(a) We want to know if this data provides strong evidence that drinking more cans of beer is associated with an increase in blood alcohol content. Write the null and alternative hypotheses in symbols, and also in words.

(b) When conducting a simulation to test the hypothesis, the p-value for this turns out to be 0. Complete the following sentence in context of the problem: "If we assume _____, then the probability of ____ is ____._"

(c) Write a conclusion in context. i.e. write a conclusion like "we have strong evidence that..." or "it is plausible that..."

(d) Does this data prove that drinking more cans of beer causes a higher BAC? Explain.

5) The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.

(a) We want to know if this data provides strong evidence that a larger height is associated with a larger weight. Write the null and alternative hypotheses in symbols, and also in words.

(b) When conducting a simulation to test the hypothesis, the p-value for this turns out to be 0. Complete the following sentence in context of the problem: "If we assume ______, then the probability of ___ is ____._"

(c) Write a conclusion in context. i.e. write a conclusion like "we have strong evidence that..." or "it is plausible that..."

(d) Would you feel confident using the regression equation to get a fairly accurate estimate of anyone’s weight based on their height? Explain.

6) A scatterplot displaying the relationship between husbands’ and wives’ ages in a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years. Given below is the scatterplot.

(a) We want to know if this data provides strong evidence that higher husband age is associated with higher wife age. Write the null and alternative hypotheses in symbols, and also in words.

(b) When conducting a simulation to test the hypothesis, the p-value for this turns out to be 0. Complete the following sentence in context of the problem: "If we assume ______, then the probability of ___ is ____._"

(c) Write a conclusion in context. i.e. write a conclusion like "we have strong evidence that..." or "it is plausible that..."

(d) Would you feel confident using the regression equation to get a fairly accurate estimate of any wife’s age based on the husband’s age? Explain.
7) The scatterplot below summarizes husbands’ and wives’ heights in a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years.

(a) We want to know if this data provides strong evidence that, for people under the age of 65, a larger height in the husband is associated with a larger height in the wife. Write the null and alternative hypotheses in symbols, and also in words.

(b) When conducting a simulation to test the hypothesis, the p-value for this turns out to be 0. Complete the following sentence in context of the problem: "If we assume __, then the probability of__ is ____.

(c) Write a conclusion in context. I.e. write a conclusion like "we have strong evidence that.." or "it is plausible that..."

(d) If you had to make an estimate of the sample correlation coefficient for this scatterplot, what would your estimate be? Explain how it is possible to get a p-value that is 0 despite having a sample correlation coefficient that is so different from 0.

8) When testing the significance of the correlation coefficient, what is the null hypothesis?

9) When testing the significance of the correlation coefficient, what is the alternative hypothesis?

10) If the level of significance is 0.05 and the p-value is 0.04, what conclusion should you draw when testing for correlation?
11) The table to the right contains real data for the first two decades of flu cases reporting.

(a) Use the Rossman Chance applet to find the sample correlation coefficient.

(b) What does the correlation coefficient imply about the strength and direction of the relationship between the year and the number of diagnosed flu cases reported in the U.S.?

(c) We want to know if this data provides strong evidence of correlation between the year and the number of flu cases diagnosed. Write the null and alternative hypotheses in symbols, and also in words.

(d) Conduct a simulation in the Rossman Chance applet to find a p-value. Complete the following sentence in context of the problem: "If we assume ___, then the probability of ____ is ___.

(e) Write a conclusion in context. I.e. write a conclusion like "we have strong evidence that.." or "it is plausible that..."

(f) Use the applet to find the t-score and p-value with a theory-based approach. Do we make the same conclusion that was made in part (e)?

(g) Would you feel confident using the regression equation to get an estimate of the number of diagnosed flu cases in any year? Explain.

<table>
<thead>
<tr>
<th>Year</th>
<th># flu cases diagnosed</th>
<th># flu deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1981</td>
<td>91</td>
<td>29</td>
</tr>
<tr>
<td>1981</td>
<td>319</td>
<td>121</td>
</tr>
<tr>
<td>1982</td>
<td>1,170</td>
<td>453</td>
</tr>
<tr>
<td>1983</td>
<td>3,076</td>
<td>1,482</td>
</tr>
<tr>
<td>1984</td>
<td>6,240</td>
<td>3,466</td>
</tr>
<tr>
<td>1985</td>
<td>11,776</td>
<td>6,878</td>
</tr>
<tr>
<td>1986</td>
<td>19,032</td>
<td>11,987</td>
</tr>
<tr>
<td>1987</td>
<td>28,564</td>
<td>16,162</td>
</tr>
<tr>
<td>1988</td>
<td>35,447</td>
<td>20,868</td>
</tr>
<tr>
<td>1989</td>
<td>42,674</td>
<td>27,591</td>
</tr>
<tr>
<td>1990</td>
<td>48,634</td>
<td>31,335</td>
</tr>
<tr>
<td>1991</td>
<td>59,660</td>
<td>36,560</td>
</tr>
<tr>
<td>1992</td>
<td>78,530</td>
<td>41,055</td>
</tr>
<tr>
<td>1993</td>
<td>78,834</td>
<td>44,730</td>
</tr>
<tr>
<td>1994</td>
<td>71,874</td>
<td>49,095</td>
</tr>
<tr>
<td>1995</td>
<td>68,505</td>
<td>49,456</td>
</tr>
<tr>
<td>1996</td>
<td>59,347</td>
<td>38,510</td>
</tr>
<tr>
<td>1997</td>
<td>47,149</td>
<td>20,736</td>
</tr>
<tr>
<td>1998</td>
<td>38,393</td>
<td>19,005</td>
</tr>
<tr>
<td>1999</td>
<td>25,174</td>
<td>18,454</td>
</tr>
<tr>
<td>2000</td>
<td>25,522</td>
<td>17,347</td>
</tr>
<tr>
<td>2001</td>
<td>25,643</td>
<td>17,402</td>
</tr>
<tr>
<td>2002</td>
<td>26,464</td>
<td>16,371</td>
</tr>
<tr>
<td>Total</td>
<td>802,118</td>
<td>489,093</td>
</tr>
</tbody>
</table>
12) Does the higher cost of tuition translate into higher-paying jobs? The table below lists the top ten colleges based on mid-career salary and the associated yearly tuition costs.

(a) Construct a scatter plot of the data.

(b) Test for correlation between mid-career salary and tuition, if the level of significance is 0.05 including a conclusion in context.

<table>
<thead>
<tr>
<th>School</th>
<th>Mid-Career Salary (in thousands)</th>
<th>Yearly Tuition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Princeton</td>
<td>137</td>
<td>28,540</td>
</tr>
<tr>
<td>Harvey Mudd</td>
<td>135</td>
<td>40,133</td>
</tr>
<tr>
<td>CalTech</td>
<td>127</td>
<td>39,900</td>
</tr>
<tr>
<td>US Naval Academy</td>
<td>122</td>
<td>0</td>
</tr>
<tr>
<td>West Point</td>
<td>120</td>
<td>0</td>
</tr>
<tr>
<td>MIT</td>
<td>118</td>
<td>42,050</td>
</tr>
<tr>
<td>Lehigh University</td>
<td>118</td>
<td>43,220</td>
</tr>
<tr>
<td>NYU-Poly</td>
<td>117</td>
<td>39,565</td>
</tr>
<tr>
<td>Babson College</td>
<td>117</td>
<td>40,400</td>
</tr>
<tr>
<td>Stanford</td>
<td>114</td>
<td>54,506</td>
</tr>
</tbody>
</table>

13) Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

<table>
<thead>
<tr>
<th>Age</th>
<th>Number of Driver Deaths per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>16–19</td>
<td>38</td>
</tr>
<tr>
<td>20–24</td>
<td>36</td>
</tr>
<tr>
<td>25–34</td>
<td>24</td>
</tr>
<tr>
<td>35–54</td>
<td>20</td>
</tr>
<tr>
<td>55–74</td>
<td>18</td>
</tr>
</tbody>
</table>

(a) Use the Rossman Chance applet to find the sample correlation coefficient. Use the mid-point of each age group as the input value that represents the age group.

(b) What does the correlation coefficient indicate about the strength and direction of the relationship between age and driver fatalities?

(c) We want to know if this data provides strong evidence of correlation between the age of the driver and number rate of driver fatalities. Write the null and alternative hypotheses in symbols, and also in words.

(d) Conduct a simulation in the Rossman Chance applet to find a p-value. Complete the following sentence in context of the problem: "If we assume_______, then the probability of ________ is ________:"

(e) Write a conclusion in context. I.e. write a conclusion like "we have strong evidence that.." or "it is plausible that..."

(f) Use the applet to find the t-score and p-value using a theory-based approach. Do we make the same conclusion that was made in part (e)?

(g) Would you feel confident using the regression equation to get an estimate of the rate of driver deaths for a given age? Explain.
14) The table below shows the life expectancy for an individual born in the United States in certain years.

<table>
<thead>
<tr>
<th>Year of Birth</th>
<th>Life Expectancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1930</td>
<td>59.7</td>
</tr>
<tr>
<td>1940</td>
<td>62.9</td>
</tr>
<tr>
<td>1950</td>
<td>70.2</td>
</tr>
<tr>
<td>1965</td>
<td>69.7</td>
</tr>
<tr>
<td>1973</td>
<td>71.4</td>
</tr>
<tr>
<td>1982</td>
<td>74.5</td>
</tr>
<tr>
<td>1987</td>
<td>75</td>
</tr>
<tr>
<td>1992</td>
<td>75.7</td>
</tr>
<tr>
<td>2010</td>
<td>78.7</td>
</tr>
</tbody>
</table>

(a) Use the Rossman Chance applet to find the sample correlation coefficient.

(b) What does the correlation coefficient indicate about the strength and direction of the relationship between year of birth and life expectancy?

(c) We want to know if this data provides strong evidence of correlation between the year of birth and life expectancy. Write the null and alternative hypotheses in symbols, and also in words.

(d) Conduct a simulation in the Rossman Chance applet to find a p-value. Complete the following sentence in context of the problem: "If we assume ________________, then the probability of ______ is _____________."

(e) Write a conclusion in context. I.e. write a conclusion like "we have strong evidence that.." or "it is plausible that..."

(f) Use the applet to find the t-score and p-value using a theory-based approach. Do we make the same conclusion that was made in part (e)?

(g) Would you feel confident using the regression equation to get an estimate of the life expectancy of a person based on their year of birth? Explain.

15) The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition ten, for various pages is given in the table below.

<table>
<thead>
<tr>
<th>Page number</th>
<th>Maximum value ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>43</td>
<td>19</td>
</tr>
<tr>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>72</td>
<td>16</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
</tr>
</tbody>
</table>

(a) Use the Rossman Chance applet to find the sample correlation coefficient.

(b) What does the correlation coefficient imply about the strength and direction of the relationship between the page number and the maximum discount value?

(c) We want to know if this data provides strong evidence of correlation between the page number and the maximum discount value. Write the null and alternative hypotheses in symbols, and also in words.

(d) Conduct a simulation in the Rossman Chance applet to find a p-value. Complete the following sentence in context of the problem: "If we assume ____, then the probability of ___________ is _____________."

(e) Write a conclusion in context. I.e. write a conclusion like "we have strong evidence that.." or "it is plausible that..."

(f) Use the applet to find the t-score and p-value using a theory-based approach. Do we make the same conclusion that was made in part (e)?

(g) Would you feel confident using the regression equation to get a fairly accurate estimate of the maximum discount value given a page number? Explain.
16) The table below gives the gold medal times for every other Summer Olympics for the women’s 100-meter freestyle (swimming).

<table>
<thead>
<tr>
<th>Year</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1912</td>
<td>82.2</td>
</tr>
<tr>
<td>1924</td>
<td>72.4</td>
</tr>
<tr>
<td>1932</td>
<td>66.8</td>
</tr>
<tr>
<td>1952</td>
<td>66.8</td>
</tr>
<tr>
<td>1960</td>
<td>61.2</td>
</tr>
<tr>
<td>1968</td>
<td>60.0</td>
</tr>
<tr>
<td>1976</td>
<td>55.65</td>
</tr>
<tr>
<td>1984</td>
<td>55.92</td>
</tr>
<tr>
<td>1992</td>
<td>54.64</td>
</tr>
<tr>
<td>2000</td>
<td>53.8</td>
</tr>
<tr>
<td>2008</td>
<td>53.1</td>
</tr>
</tbody>
</table>

(a) Use the Rossman Chance applet to find the sample correlation coefficient.

(b) What does the correlation coefficient imply about the strength and direction of the relationship between the year and the winning swimming time?

(c) We want to know if this data provides strong evidence of correlation between the year and the winning swimming time. Write the null and alternative hypotheses in symbols, and also in words.

(d) Conduct a simulation in the Rossman Chance applet to find a p-value. Complete the following sentence in context of the problem: "If we assume __________, then the probability of__________ is__________.

(e) Write a conclusion in context. i.e. write a conclusion like "we have strong evidence that.." or "it is plausible that..."

(f) Use the applet to find the t-score and p-value with a theory-based approach. Do we make the same conclusion that was made in part (e)?

(g) Would you feel confident using the regression equation to get a fairly accurate estimate of the winning swimming time given the year? Explain.

17) We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union. Conduct a hypothesis test using simulation in the Rossman Chance applet, and write a conclusion in context based on the simulated p-value. Be sure to clearly describe and detail all steps of the hypothesis-testing process.

<table>
<thead>
<tr>
<th>State</th>
<th># letters in name</th>
<th>Year entered the Union</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>7</td>
<td>1819</td>
</tr>
<tr>
<td>Colorado</td>
<td>8</td>
<td>1876</td>
</tr>
<tr>
<td>Hawaii</td>
<td>6</td>
<td>1959</td>
</tr>
<tr>
<td>Iowa</td>
<td>4</td>
<td>1846</td>
</tr>
<tr>
<td>Maryland</td>
<td>8</td>
<td>1788</td>
</tr>
<tr>
<td>Missouri</td>
<td>8</td>
<td>1821</td>
</tr>
<tr>
<td>New Jersey</td>
<td>9</td>
<td>1787</td>
</tr>
<tr>
<td>Ohio</td>
<td>4</td>
<td>1803</td>
</tr>
<tr>
<td>South Carolina</td>
<td>13</td>
<td>1788</td>
</tr>
<tr>
<td>Utah</td>
<td>4</td>
<td>1896</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>9</td>
<td>1848</td>
</tr>
</tbody>
</table>