Advanced High School Statistics

Second Edition,
with updates based on AP® Statistics Course Framework

David Diez
Data Scientist
OpenIntro

Mine Çetinkaya-Rundel
Associate Professor of the Practice, Duke University
Professional Educator, RStudio

Leah Dorazio
Statistics and Computer Science Teacher
San Francisco University High School

Christopher D Barr
Investment Analyst
Varadero Capital

Table of Contents

Preface

Advanced High School Statistics covers a first course in statistics, providing an introduction to applied statistics that is clear, concise, and accessible. This book was written to align with the AP® Statistics Course Description[1], but it's also popular in non-AP courses and community colleges.

This book may be downloaded as a free PDF at **openintro.org/ahss**.

We hope readers will take away three ideas from this book in addition to forming a foundation of statistical thinking and methods.

(1) Statistics is an applied field with a wide range of practical applications.

(2) You don't have to be a math guru to learn from real, interesting data.

(3) Data are messy, and statistical tools are imperfect. But, when you understand the strengths and weaknesses of these tools, you can use them to learn about the real world.

Textbook overview

The chapters of this book are as follows:

1. **Data collection.** Data structures, variables, and basic data collection techniques.

2. **Summarizing data.** Data summaries and graphics.

3. **Probability.** The basic principles of probability.

4. **Distributions of random variables.** Introduction to key distributions, and how the normal model applies to the sample mean and sample proportion.

5. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population proportion.

6. **Inference for categorical data.** Inference for proportions and contingency tables using the normal and chi-square distributions.

7. **Inference for numerical data.** Inference for one or two sample means using the t-distribution.

8. **Introduction to linear regression.** An introduction to regression with two variables, and inference on the slope of the regression line.

Online resources

OpenIntro is focused on increasing access to education by developing free, high-quality education materials. In addition to textbooks, we provide the following accompanying resources to help teachers and students be successful.

- Video overviews for each section of the textbook
- Lecture slides for each section of the textbook
- Casio and TI calculator tutorials
- Video solutions for selected section and chapter exercises

---

[1]AP® is a trademark registered and owned by the College Board, which was not involved in the production of, and does not endorse, this product. apcentral.collegeboard.org/pdf/ap-statistics-course-description.pdf

- Statistical software labs
- A small but growing number of Desmos activities[2]
- Quizlet sets for each chapter[3]
- A Tableau public page to further interact with data sets[4]
- Online, interactive version of textbook[5]
- Complete companion course with the learning management software MyOpenMath[6]
- Complete Canvas course accessible through Canvas Commons[7]

All of these resources can be found at:

openintro.org/ahss

We also have improved the ability to access data in this book through the addition of Appendix B, which provides additional information for each of the data sets used in the main text and is new in the Second Edition. Online guides to each of these data sets are also provided at **openintro.org/data** and through a companion R package.

Examples and exercises

Many examples are provided to establish an understanding of how to apply methods.

E

EXAMPLE 0.1

This is an example.

―――――――――

Full solutions to examples are provided here, within the example.

When we think the reader should be ready to do an example problem on their own, we frame it as Guided Practice.

G

GUIDED PRACTICE 0.2

The reader may check or learn the answer to any Guided Practice problem by reviewing the full solution in a footnote.[8]

Exercises are also provided at the end of each section and each chapter for practice or homework assignments. Solutions for odd-numbered exercises are given in Appendix A.

Getting involved

We encourage anyone learning or teaching statistics to visit **openintro.org** and get involved. We value your feedback. Please send any questions or comments to leah@openintro.org. You can also provide feedback, report typos, and review known typos at

openintro.org/ahss/feedback

Acknowledgements

This project would not be possible without the passion and dedication of all those involved. The authors would like to thank the OpenIntro Staff for their involvement and ongoing contributions. We are also very grateful to the hundreds of students and instructors who have provided us with valuable feedback since we first started working on this project in 2009. A special thank you to Catherine Ko for proofreading the second edition of AHSS.

―――――――――――――――――――

[2] openintro.org/ahss/desmos
[3] quizlet.com/openintro-ahss
[4] public.tableau.com/profile/openintro
[5] Developed by Emiliano Vega and Ralf Youtz of Portland Community College using PreTeXt.
[6] myopenmath.com/course/public.php?cid=11774
[7] sfuhs.instructure.com/courses/1068
[8] Guided Practice solutions are always located down here!

Chapter 1

Data collection

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called **data**. Statistics is the study of how best to collect, analyze, and draw conclusions from data. It is helpful to put statistics in the context of a general process of investigation:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Researchers from a wide array of fields have questions or problems that require the collection and analysis of data. What questions from current events or from your own life can you think of that could be answered by collecting and analyzing data?

This chapter focuses on collecting data. We'll discuss basic properties of data, common sources of bias that arise during data collection, and techniques for collecting data. After finishing this chapter, you will have the tools for identifying weaknesses and strengths in data-based conclusions, tools that are essential to be an informed citizen and a savvy consumer of information.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

1.1   Case study: using stents to prevent strokes

We start with a case study and we consider the following questions:

- Does the use of stents reduce the risk of stroke?

- How do researchers collect data to answer this question?

- What do they do with the data once it is collected?

- How different must the risk of stroke be in each group before there is sufficient evidence that it's a real difference and not just random variation?

Learning objectives

1. Understand the four steps of a statistical investigation (identify a question, collect data, analyze data, form a conclusion) in the context of a real-world example.

2. Consider the concept of statistical significance.

1.1.1   Case study

Section 1.1 introduces a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke.[1] Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

Does the use of stents reduce the risk of stroke?

The researchers who asked this question collected data on 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

Treatment group. Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.

Control group. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

---

[1]Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335. NY Times article reporting on the study: www.nytimes.com/2011/09/08/health/research/08stent.html.

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Figure 1.1. Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Figure 1.1: Results for five patients from the stent study.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Figure 1.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke within 30 days, we look on the left-side of the table at the intersection of the treatment and stroke: 33.

	0-30 days			0-365 days	
	stroke	no event		stroke	no event
treatment	33	191		45	179
control	13	214		28	199
Total	46	405		73	378

Figure 1.2: Descriptive statistics for the stent study.

GUIDED PRACTICE 1.1

What proportion of the patients in the treatment group had no stroke within the first 30 days of the study? (Please note: answers to all Guided Practice exercises are provided using footnotes.)[2]

We can compute summary statistics from the table. A **summary statistic** is a single number summarizing a large amount of data.[3] For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.

Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So what we are really asking is whether the difference is **statistically significant**, that is, whether the difference so large that we should reject the notion that it was due to chance.

---

[2]There were 191 patients in the treatment group that had no stroke in the first 30 days. There were 33 + 191 = 224 total patients in the treatment group, so the proportion is 191/224 = 0.85.

[3]Formally, a summary statistic is a value computed from the data. Some summary statistics are more useful than others.

While we don't yet have the statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

Section summary

- To test the effectiveness of a treatment, researchers often carry out an experiment in which they randomly assign patients to a **treatment group** or a **control group**.

- Researchers compare the relevant **summary statistics** to get a sense of whether the treatment group did better, on average, than the control group.

- Ultimately, researchers want to know whether the difference between the two groups is **significant**, that is, larger than what would be expected by chance alone.

Exercises

1.1 Migraine and acupuncture, Part 1. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[4]

<table>
<tr><td></td><td></td><td colspan="2"><i>Pain free</i></td><td></td></tr>
<tr><td></td><td></td><td>Yes</td><td>No</td><td>Total</td></tr>
<tr><td rowspan="2"><i>Group</i></td><td>Treatment</td><td>10</td><td>33</td><td>43</td></tr>
<tr><td>Control</td><td>2</td><td>44</td><td>46</td></tr>
<tr><td></td><td>Total</td><td>12</td><td>77</td><td>89</td></tr>
</table>
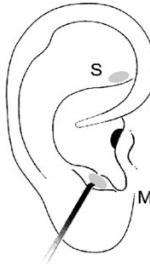
Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

(a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?

(b) What percent were pain free in the control group?

(c) In which group did a higher percent of patients become pain free 24 hours after receiving acupuncture?

(d) Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people who suffer from migraines. However this is not the only possible conclusion that can be drawn based on your findings so far. What is one other possible explanation for the observed difference between the percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

1.2 Sinusitis and antibiotics, Part 1. Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. The distribution of responses is summarized below.[5]

<table>
<tr><td></td><td></td><td colspan="2"><i>Self-reported improvement
in symptoms</i></td><td></td></tr>
<tr><td></td><td></td><td>Yes</td><td>No</td><td>Total</td></tr>
<tr><td rowspan="2"><i>Group</i></td><td>Treatment</td><td>66</td><td>19</td><td>85</td></tr>
<tr><td>Control</td><td>65</td><td>16</td><td>81</td></tr>
<tr><td></td><td>Total</td><td>131</td><td>35</td><td>166</td></tr>
</table>

(a) What percent of patients in the treatment group experienced improvement in symptoms?

(b) What percent experienced improvement in symptoms in the control group?

(c) In which group did a higher percentage of patients experience improvement in symptoms?

(d) Your findings so far might suggest a real difference in effectiveness of antibiotic and placebo treatments for improving symptoms of sinusitis. However, this is not the only possible conclusion that can be drawn based on your findings so far. What is one other possible explanation for the observed difference between the percentages of patients in the antibiotic and placebo treatment groups that experience improvement in symptoms of sinusitis?

---

[4]G. Allais et al. "Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints". In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

[5]J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

1.2 Data basics

You collect data on dozens of questions from all of the students at your school. How would you organize all of this data? Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book. We use loan data from Lending Club and county data from the US Census Bureau to motivate and illustrate this section's learning objectives.

Learning objectives

1. Identify the individuals and the variables of a study.

2. Identify variables as categorical or numerical. Identify numerical variables as discrete or continuous.

3. Understand what it means for two variables to be associated.

1.2.1 Observations, variables, and data matrices

Figure 1.3 displays rows 1, 2, 3, and 50 of a data set for 50 randomly sampled loans offered through Lending Club, which is a peer-to-peer lending company. These observations will be referred to as the `loan50` data set.

Each row in the table represents a single loan. The formal name for a row is a **case** or **observational unit**. The columns represent characteristics, called **variables**, for each of the loans. For example, the first row represents a loan of $7,500 with an interest rate of 7.34%, where the borrower is based in Maryland (MD) and has an income of $70,000.

GUIDED PRACTICE 1.2

What is the grade of the first loan in Figure 1.3? And what is the home ownership status of the borrower for that first loan? For these Guided Practice questions, you can check your answer in the footnote.[6]

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of the `loan50` variables are given in Figure 1.4.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Figure 1.3: Four rows from the `loan50` data matrix.

---

[6]The loan's grade is A, and the borrower rents their residence.

variable	description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Figure 1.4: Variables and their descriptions for the loan50 data set.

The data in Figure 1.3 represent a **data matrix**, which is a convenient and common way to organize data, especially if collecting data in a spreadsheet. Each row of a data matrix corresponds to a unique case (observational unit), and each column corresponds to a variable.

When recording data, use a data matrix unless you have a very good reason to use a different structure. This structure allows new cases to be added as rows or new variables as new columns.

GUIDED PRACTICE 1.3

The grades for assignments, quizzes, and exams in a course are often recorded in a gradebook that takes the form of a data matrix. How might you organize grade data using a data matrix?[7]

GUIDED PRACTICE 1.4

We consider data for 3,142 counties in the United States, which includes each county's name, the state in which it is located, its population in 2017, how its population changed from 2010 to 2017, poverty rate, and six additional characteristics. How might these data be organized in a data matrix?[8]

The data described in Guided Practice 1.4 represents the county data set, which is shown as a data matrix in Figure 1.5. These data come from the US Census, with much of the data coming from the US Census Bureau's American Community Survey (ACS). Unlike the Decennial Census, which takes place every 10 years and attempts to collect basic demographic data from every residents of the US, the ACS is an ongoing survey that is sent to approximately 3.5 million households per year. As stated by the ACS website, these data help communities "plan for hospitals and schools, support school lunch programs, improve emergency services, build bridges, and inform businesses looking to add jobs and expand to new markets, and more."[9] A small subset of the variables from the ACS are summarized in Figure 1.6.

---

[7]There are multiple strategies that can be followed. One common strategy is to have each student represented by a row, and then add a column for each assignment, quiz, or exam. Under this setup, it is easy to review a single line to understand a student's grade history. There should also be columns to include student information, such as one column to list student names.

[8]Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,142 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

[9]https://www.census.gov/programs-surveys/acs/about.html

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	hs_diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	yes	hs_diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	hs_diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	hs_diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	hs_diploma	40041
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

Figure 1.5: Eleven rows from the county data set.

variable	description
name	County name.
state	State where the county resides, or the District of Columbia.
pop	Population in 2017.
pop_change	Percent change in the population from 2010 to 2017. For example, the value 1.48 in the first row means the population for this county increased by 1.48% from 2010 to 2017.
poverty	Percent of the population in poverty.
homeownership	Percent of the population that lives in their own home or lives with the owner, e.g. children living with parents who own the home.
multi_unit	Percent of living units that are in multi-unit structures, e.g. apartments.
unemp_rate	Unemployment rate as a percent.
metro	Whether the county contains a metropolitan area.
median_edu	Median education level, which can take a value among below_hs, hs_diploma, some_college, and bachelors.
median_hh_income	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older.

Figure 1.6: Variables and their descriptions for the county data set.

1.2.2 Types of variables

Examine the `unemp_rate`, `pop`, `state`, and `median_edu` variables in the `county` data set. Each of these variables is inherently different from the other three, yet some share certain characteristics.

First consider `unemp_rate`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since the average, sum, and difference of area codes doesn't have any clear meaning.

The `pop` variable is also numerical, although it seems to be a little different than `unemp_rate`. This variable of the population count can only take whole non-negative numbers (0, 1, 2, ...). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the unemployment rate variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: `AL`, `AK`, ..., and `WY`. Because the responses themselves are categories, `state` is called a **categorical** variable, and the possible values are called the variable's **levels**.

Finally, consider the `median_edu` variable, which describes the median education level of county residents and takes values `below_hs`, `hs_diploma`, `some_college`, or `bachelors` in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable, while a regular categorical variable without this type of special ordering is called a **nominal** variable. To simplify analyses, any ordinal variable in this book will be treated as a nominal (unordered) categorical variable.
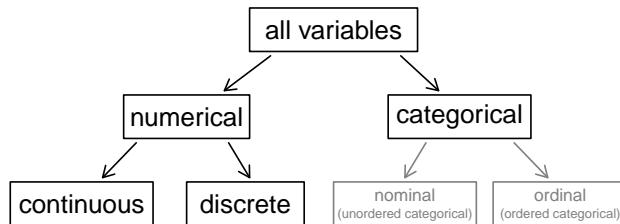
Figure 1.7: Breakdown of variables into their respective types.

EXAMPLE 1.5

Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

———

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

GUIDED PRACTICE 1.6

An experiment is evaluating the effectiveness of a new drug in treating migraines. A `group` variable is used to indicate the experiment group for each patient: treatment or control. The `num_migraines` variable represents the number of migraines the patient experienced during a 3-month period. Classify each variable as either numerical or categorical.[10]

———

[10]The `group` variable can take just one of two group names, making it categorical. The `num_migraines` variable describes a count of the number of migraines, which is an outcome where basic arithmetic is sensible, which means this is a numerical outcome; more specifically, since it represents a count, `num_migraines` is a discrete numerical variable.

1.2.3   Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

(1) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county tend to be above or below the national average?

(2) Does a higher than average increase in county population tend to correspond to counties with higher or lower median household incomes?

(3) How useful a predictor is median education level for the median household income for US counties?

To answer these questions, data must be collected, such as the `county` data set shown in Figure 1.5. Examining summary statistics could provide insights for each of the three questions about counties. Additionally, graphs can be used to visually explore the data.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables `homeownership` and `multi_unit`, which is the percent of units in multi-unit structures (e.g. apartments, condos). Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 413 in the `county` data set: Chattahoochee County, Georgia, which has 39.4% of units in multi-unit structures and a homeownership rate of 31.3%. The scatterplot suggests a relationship between the two variables: counties with a higher rate of multi-units tend to have lower homeownership rates. We might brainstorm as to why this relationship exists and investigate the ideas to determine which are the most reasonable explanations.
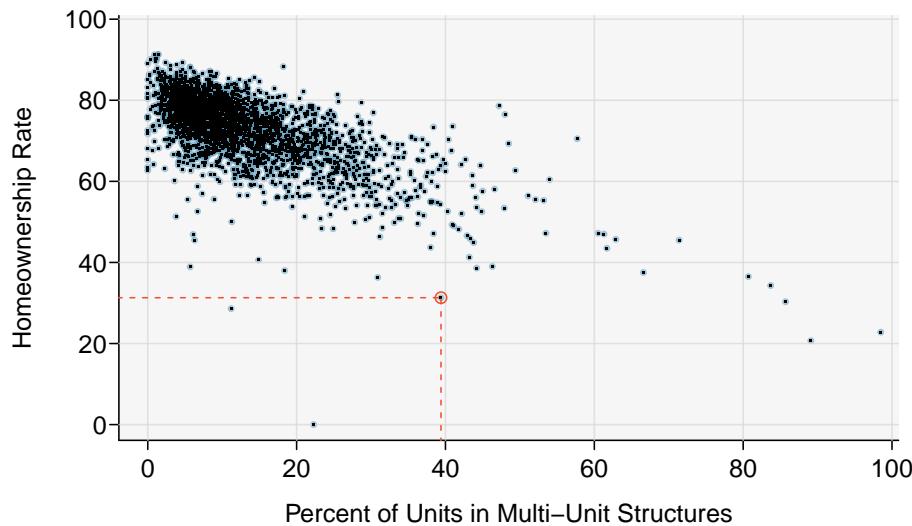
Figure 1.8: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for US counties. The highlighted dot represents Chattahoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeownership rate of 31.3%. Explore this scatterplot and dozens of other scatterplots using American Community Survey data on Tableau Public⁺↖.

The multi-unit and homeownership rates are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables. Associated variables can also be called **dependent** variables and vice-versa.
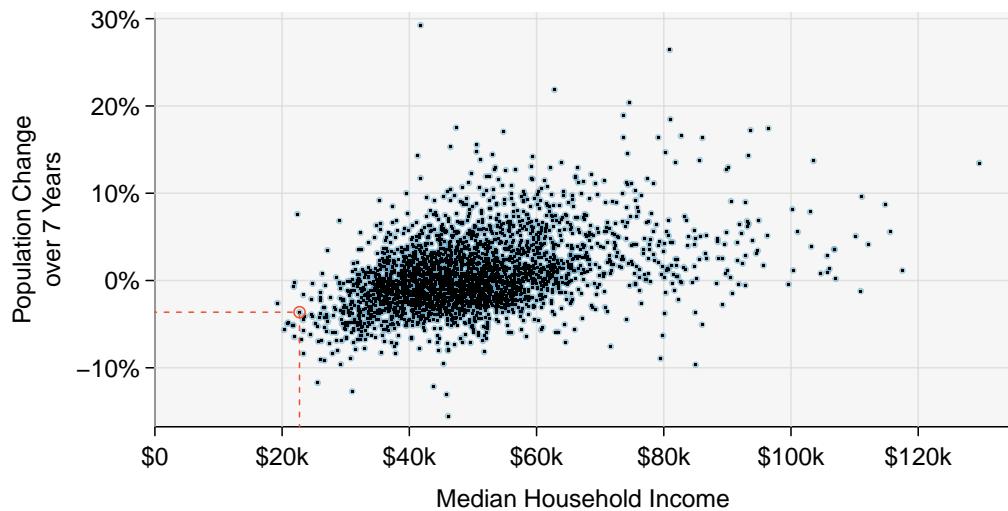
Figure 1.9: A scatterplot showing `pop_change` against `median_hh_income`. Owsley County of Kentucky, is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of $22,736. Explore this scatterplot and dozens of other scatterplots using American Community Survey data on Tableau Public⁺↖.

GUIDED PRACTICE 1.7

Examine the variables in the `loan50` data set, which are described in Figure 1.4 on page 18. Create two questions about possible relationships between variables in `loan50` that are of interest to you.[11]

EXAMPLE 1.8

This example examines the relationship between a county's population change from 2010 to 2017 and median household income, which is visualized as a scatterplot in Figure 1.9. Are these variables associated?

The larger the median household income for a county, the higher the population growth observed for the county. While this trend isn't true for every county, the trend in the plot is evident. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.8 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `median_hh_income` and `pop_change` in Figure 1.9, where counties with higher median household income tend to have higher rates of population growth.

If two variables are not associated, then they are said to be **independent**. That is, two variables are independent if there is no evident relationship between the two.

ASSOCIATED OR INDEPENDENT, NOT BOTH

A pair of variables is either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

---

[11]Two example questions: (1) What is the relationship between loan amount and total income? (2) If someone's income is above the average, will their interest rate tend to be above or below the average?

Section summary

- Researchers often summarize data in a table, where the rows correspond to individuals or **cases** and the columns correspond to the **variables**, the values of which are recorded for each individual.

- Variables can be **numerical** (measured on a numerical scale) or **categorical** (taking on levels, such as low/medium/high). Numerical variables can be **continuous**, where all values within a range are possible, or **discrete**, where only specific values, usually integer values, are possible.

- When there exists a relationship between two variables, the variables are said to be **associated** or **dependent**. If the variables are not associated, they are said to be **independent**.

Exercises

1.3   Air pollution and birth outcomes, study components. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[12]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.4   Buteyko method, study components. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomnly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[13]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.5   Cheaters, study components. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white.[14]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) The study's findings can be summarized as follows: "Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls." How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

---

[12]B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502–511.

[13]J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

[14]Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". In: *Journal of Economic Psychology* 32.1 (2011), pp. 73–78.

1.6 Stealers, study components. In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken.[15]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

1.7 Migraine and acupuncture, Part 2. Exercise 1.1 introduced a study exploring whether acupuncture had any effect on migraines. Researchers conducted a randomized controlled study where patients were randomly assigned to one of two groups: treatment or control. The patients in the treatment group received acupuncture that was specifically designed to treat migraines. The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. What are the explanatory and response variables in this study?

1.8 Sinusitis and antibiotics, Part 2. Exercise 1.2 introduced a study exploring the effect of antibiotic treatment for acute sinusitis. Study participants either received either a 10-day course of an antibiotic (treatment) or a placebo similar in appearance and taste (control). At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. What are the explanatory and response variables in this study?

1.9 Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.[16]

(a) How many cases were included in the data?

(b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

(c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Photo by Ryan Claussen (http://flic.kr/p/6QTcuX) CC BY-SA 2.0 license

1.10 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[17]

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day
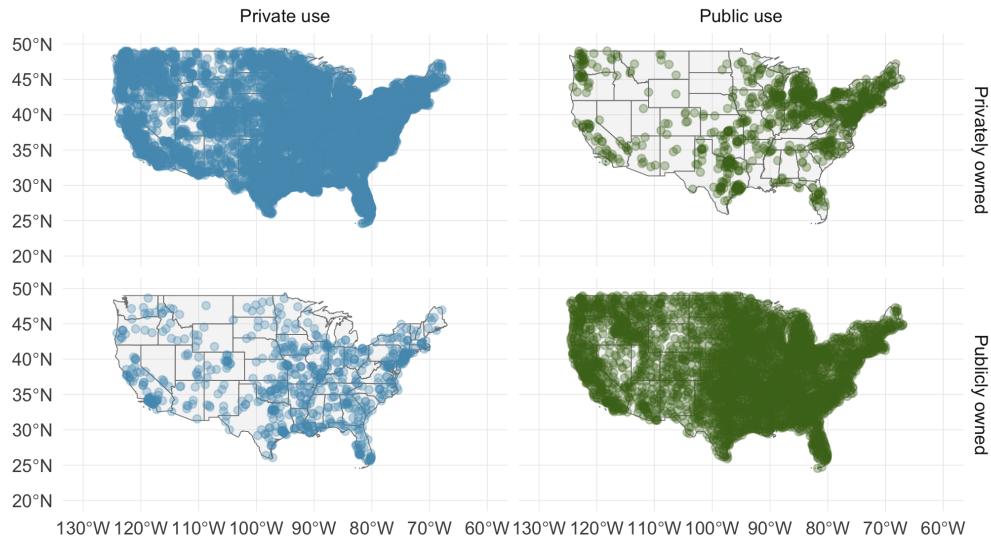
(a) What does each row of the data matrix represent?

(b) How many participants were included in the survey?

(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

---

[15]P.K. Piff et al. "Higher social class predicts increased unethical behavior". In: *Proceedings of the National Academy of Sciences* (2012).

[16]R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.
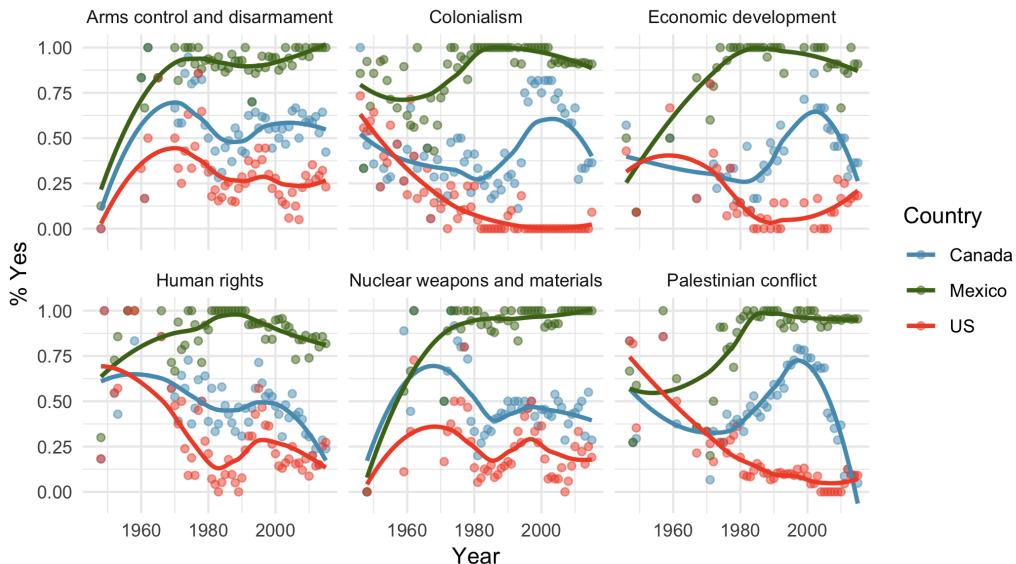
[17]National STEM Centre, Large Datasets from stats4schools.

1.11   US Airports. The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.[18]

(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

1.12   UN Votes. The visualization below shows voting patterns the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.[19]

(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

---

[18]Federal Aviation Administration, www.faa.gov/airports/airport_safety/airportdata_5010.

[19]David Robinson. *unvotes: United Nations General Assembly Voting Data*. R package version 0.2.0. 2017. URL: https://CRAN.R-project.org/package=unvotes.

1.3   Overview of data collection principles

How do researchers collect data? Why are the results of some studies more reliable than others? The way a researcher collects data depends upon the research goals. In this section, we look at different methods of collecting data and consider the types of conclusions that can be drawn from those methods.

Learning objectives

1. Distinguish between the population and a sample and between the parameter and a statistic.

2. Know when to summarize a data set using a mean versus a proportion.

3. Understand why anecdotal evidence is unreliable.

4. Identify the four main types of data collection: census, sample survey, experiment, and observation study.

5. Classify a study as observational or experimental, and determine when a study's results can be generalized to the population and when a causal relationship can be drawn.

1.3.1   Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?

2. Over the last 5 years, what is the average time to complete a degree for Duke undergrads?

3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A **sample** represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

G

GUIDED PRACTICE 1.9

For the second and third questions above, identify the target population and what represents an individual case.[20]

---

[20](2) Notice that this question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergrads who have graduated in the last five years are part of the population of interest. Each such student would represent an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

We collect a sample of data to better understand the characteristics of a population. A **variable** is a characteristic we measure for each individual or case. The overall quantity of interest may be the mean, median, proportion, or some other summary of a population. These population values are called **parameters**. We estimate the value of a parameter by taking a sample and computing a numerical summary called a **statistic** based on that sample. Note that the two p's (population, parameter) go together and the two s's (sample, statistic) go together.

EXAMPLE 1.10

(E)

Earlier we asked the question: what is the average mercury content in swordfish in the Atlantic Ocean? Identify the variable to be measured and the parameter and statistic of interest.

—————

The variable is the level of mercury content in swordfish in the Atlantic Ocean. It will be measured for each individual swordfish. The parameter of interest is the average mercury content in *all* swordfish in the Atlantic Ocean. If we take a sample of 50 swordfish from the Atlantic Ocean, the average mercury content among just those 50 swordfish will be the statistic.

Two statistics we will study are the **mean** (also called the **average**) and **proportion**. When we are discussing a population, we label the mean as μ (the Greek letter, mu), while we label the sample mean as $\bar{x}$ (read as *x-bar*). When we are discussing a proportion in the context of a population, we use the label p, while the sample proportion has a label of $\hat{p}$ (read as *p-hat*). Generally, we use $\bar{x}$ to estimate the population mean, μ. Likewise, we use the sample proportion $\hat{p}$ to estimate the population proportion, p.

EXAMPLE 1.11

(E)

Is μ a parameter or statistic? What about $\hat{p}$?

—————

μ is a parameter because it refers to the average of the *entire* population. $\hat{p}$ is a statistic because it is calculated from a sample.

EXAMPLE 1.12

(E)

For the second question regarding time to complete a degree for a Duke undergraduate, is the variable numerical or categorical? What is the parameter of interest?

—————

The characteristic that we record on each individual is the number of years until graduation, which is a numerical variable. The parameter of interest is the average time to degree for all Duke undergraduates, and we use μ to describe this quantity.

GUIDED PRACTICE 1.13

(G)

The third question asked whether a new drug reduces deaths in patients with severe heart disease. Is the variable numerical or categorical? Describe the statistic that should be calculated in this study.[21]

If these topics are still a bit unclear, don't worry. We'll cover them in greater detail in the next chapter.

—————————————

[21]The variable is whether or not a patient with severe heart disease dies within the time frame of the study. This is categorical because it will be a yes or a no. The statistic that should be recorded is the proportion of patients that die within the time frame of the study, and we would use $\hat{p}$ to denote this quantity.

Figure 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, "It's one storm, in one region, of one country."

———————————

February 10th, 2010.

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.

2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.

3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

ANECDOTAL EVIDENCE

Be careful of making inferences based on anecdotal evidence. Such evidence may be true and verifiable, but it may only represent extraordinary cases. The majority of cases and the average case may in fact be very different.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we may vividly remember the time when our friend bought a lottery ticket and won $250 but forget most the times she bought one and lost. Instead of focusing on the most unusual cases, we should examine a representative sample of many cases.

1.3.3 Explanatory and response variables
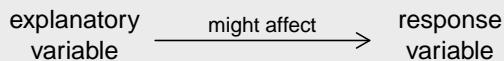
When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other. Consider the following rephrasing of an earlier question about the `county` data set:

> *If there is an increase in the median household income in a county, does this drive an increase in its population?*

In this question, we are asking whether one variable affects another. If this is our underlying belief, then *median household income* is the **explanatory** variable and the *population change* is the **response** variable in the hypothesized relationship.[22]

EXPLANATORY AND RESPONSE VARIABLES

When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable.

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable}$$

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

ASSOCIATION DOES NOT IMPLY CAUSATION

Labeling variables as *explanatory* and *response* does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

In many cases, the relationship is complex or unknown. It may be unclear whether variable A explains variable B or whether variable B explains variable A. For example, it is now known that a particular protein called REST is much depleted in people suffering from Alzheimer's disease. While this raises hopes of a possible approach for treating Alzheimer's, it is still unknown whether the lack of the protein causes brain deterioration, whether brain deterioration causes depletion in the REST protein, or whether some third variable causes both brain deterioration and REST depletion. That is, we do not know if the lack of the protein is an explanatory variable or a response variable. Perhaps it is both.[23]

---

[22]Sometimes the explanatory variable is called the **independent** variable and the response variable is called the **dependent** variable. However, this becomes confusing since a *pair* of variables might be independent or dependent, so we avoid this language.

[23]nytimes.com/2014/03/20/health/fetal-gene-may-protect-brain-from-alzheimers-study-finds.html

1.3.4 Observational studies versus experiments

There are two primary types of data collection: observational studies and experiments.

Researchers perform an **observational study** when they collect data without interfering with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to study why certain diseases might develop. In each of these situations, researchers merely observe or take measurements of things that arise naturally.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. For all experiments, the researchers must impose a treatment. For most studies there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. For example, each heart attack patient in the drug trial could be randomly assigned into one of two groups: the first group receives a **placebo** (fake treatment) and the second group receives the drug. See the case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

EXAMPLE 1.14

Suppose that a researcher is interested in the average tip customers at a particular restaurant give. Should she carry out an observational study or an experiment?

In addressing this question, we ask, "Will the researcher be imposing any treatment?" Because there is no treatment or interference that would be applicable here, it will be an observational study. Additionally, one consideration the researcher should be aware of is that, if customers know their tips are being recorded, it could change their behavior, making the results of the study inaccurate.

ASSOCIATION ≠ CAUSATION

In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

Section summary

- The **population** is the entire group that the researchers are interested in. Because it is usually too costly to gather the data for the entire population, researchers will collect data from a **sample**, representing a subset of the population.

- A **parameter** is a true quantity for the entire population, while a **statistic** is what is calculated from the sample. A parameter is about a population and a statistic is about a sample. Remember: *p goes with p and s goes with s*.

- Two common summary quantities are **mean** (for numerical variables) and **proportion** (for categorical variables).

- Finding a good estimate for a population parameter requires a random sample; do not generalize from anecdotal evidence.

- There are two primary types of data collection: observational studies and experiments. In an **experiment**, researchers impose a treatment to look for a causal relationship between the treatment and the response. In an **observational study**, researchers simply collect data without imposing any treatment.

- Remember: *Correlation is not causation*! In other words, an association between two variables does not imply that one causes the other. Proving a causal relationship requires a well-designed experiment.

Exercises

1.13   Air pollution and birth outcomes, scope of inference.   Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.14   Cheaters, scope of inference.   Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.15   Buteyko method, scope of inference.   Exercise 1.4 introduces a study on using the Buteyko shallow breathing technique to reduce asthma symptoms and improve quality of life. As part of this study 600 asthma patients aged 18-69 who relied on medication for asthma treatment were recruited and randomly assigned to two groups: one practiced the Buteyko method and the other did not. Those in the Buteyko group experienced, on average, a significant reduction in asthma symptoms and an improvement in quality of life.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.16   Stealers, scope of inference.   Exercise 1.6 introduces a study on the relationship between socioeconomic class and unethical behavior. As part of this study 129 University of California Berkeley undergraduates were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.17   Relaxing after work.   The General Social Survey asked the question, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

(a) An American in the sample.

(b) Number of hours spent relaxing after an average work day.

(c) 1.65.

(d) Average number of hours all Americans spend relaxing after an average work day.

1.18 Cats on YouTube. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos.Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

(a) Percentage of all videos on YouTube that are cat videos.

(b) 2%.

(c) A video in your sample.

(d) Whether or not a video is a cat video.

1.4 Observational studies and sampling strategies

You have probably read or heard claims from many studies and polls. A background in statistical reasoning will help you assess the validity of such claims. Some of the big questions we address in this section include:

- If a study finds a relationship between two variables, such as eating chocolate and positive health outcomes, is it reasonable to conclude eating chocolate improves health outcomes?
- How do opinion polls work? How do research organizations collect the data, and what types of bias should we look out for?

Learning objectives

1. Identify possible confounding factors in a study and explain, in context, how they could confound.
2. Distinguish among and describe a convenience sample, a volunteer sample, and a random sample.
3. Identify and describe the effects of different types of bias in sample surveys, including undercoverage, non-response, and response bias.
4. Identify and describe how to implement different random sampling methods, including simple, systematic, stratified, and cluster.
5. Recognize the benefits and drawbacks of choosing one sampling method over another.
6. Understand when it valid to generalize and to what population that generalization can be made.

1.4.1 Observational studies

Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.
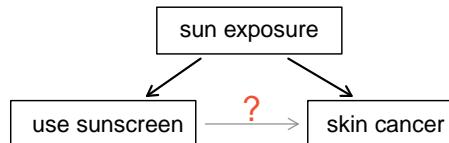
Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data is treacherous and is not recommended. Observational studies are generally only sufficient to show associations.

GUIDED PRACTICE 1.15

Ⓖ Suppose an observational study tracked sunscreen use and skin cancer, and it was found people who use sunscreen are more likely to get skin cancer than people who do not use sunscreen. Does this mean sunscreen *causes* skin cancer?[24]

---

[24]No. See the paragraph following the exercise for an explanation.

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. Sun exposure is what is called a **confounding variable** (also called a **lurking variable**, **confounding factor**, or a **confounder**).

CONFOUNDING VARIABLE

A confounding variable is a variable that is associated with both the explanatory *and* response variables. Because of the confounding variable's association with both variables, we do not know if the response is due to the explanatory variable or due to the confounding variable.

Sun exposure is a confounding factor because it is associated with both the use of sunscreen and the development of skin cancer. People who are out in the sun all day are more likely to use sunscreen, and people who are out in the sun all day are more likely to get skin cancer. Research shows us the development of skin cancer is due to the sun exposure. The variables of sunscreen usage and sun exposure are **confounded**, and without this research, we would have no way of knowing which one was the true cause of skin cancer.

EXAMPLE 1.16

In a study that followed 1,169 non-diabetic men and women who had been hospitalized for a first heart attack, the people that reported eating chocolate had increased survival rate over the next 8 years than those that reported not eating chocolate.[25] Also, those who ate more chocolate also tended to live longer on average. The researched controlled for several confounding factors, such as age, physical activity, smoking, and many other factors. Can we conclude that the consumption of chocolate caused the people to live longer?

This is an observational study, not a controlled randomized experiment. Even though the researchers controlled for many possible variables, there may still be other confounding factors. (Can you think of any that weren't mentioned?) While it is possible that the chocolate had an effect, this study cannot prove that chocolate increased the survival rate of patients.

EXAMPLE 1.17

The authors who conducted the study did warn in the article that additional studies would be necessary to determine whether the correlation between chocolate consumption and survival translates to any causal relationship. That is, they acknowledged that there may be confounding factors. One possible confounding factor not considered was mental health. In context, explain what it would mean for mental health to be a confounding factor in this study.

Mental health would be a confounding factor if, for example, people with better mental health tended to eat more chocolate, and those with better mental health *also* were less likely to die within the 8 year study period. Notice that if better mental health were not associated with eating more chocolate, it would not be considered a confounding factor since it wouldn't explain the observed associated between eating chocolate and having a better survival rate. If better mental health were associated only with eating chocolate and not with a better survival rate, then it would also not be confounding for the same reason. Only if a variable that is associated with both the explanatory variable of interest (chocolate) and the outcome variable in the study (survival during the 8 year study period) can it be considered a confounding factor.

---

[25]Janszky et al. 2009. Chocolate consumption and mortality following a first acute myocardial infarction: the Stockholm Heart Epidemiology Program. Journal of Internal Medicine 266:3, p248-257.

While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

In the same way, the `county` data set is an observational study with confounding variables, and its data cannot be used to make causal conclusions.

GUIDED PRACTICE 1.18

(G)

Figure 1.8 shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest one or more other variables that might explain the relationship visible in Figure 1.8.[26]

Observational studies come in two forms: prospective and retrospective studies. A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of similar individuals over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses' Health Study, started in 1976 and expanded in 1989.[27] This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets, such as `county`, may contain both prospectively- and retrospectively-collected variables. Local governments prospectively collect some variables as events unfolded (e.g. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

1.4.2 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. The goal is to use information from the sample to generalize or make an inference to the population. In order to be able to generalize, we must *randomly* select a sample from the population of interest. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates.
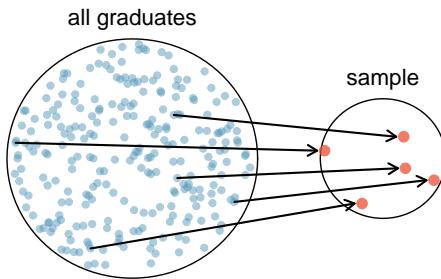
Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

Why pick a sample randomly? Why not just pick a sample by hand? Consider the following scenario.

---

[26]Answers will vary. Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.

[27]www.channing.harvard.edu/nhs
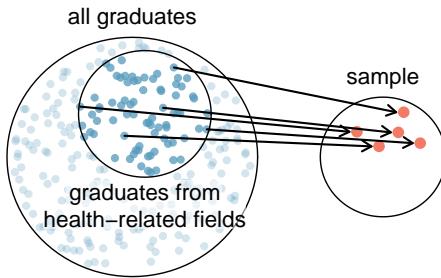
Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

EXAMPLE 1.19

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might select? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population. When selecting samples by hand, we run the risk of picking a *biased* sample, even if that bias is unintentional or difficult to discern.

If the student majoring in nutrition picked a disproportionate number of graduates from health-related fields, this would introduce undercoverage bias into the sample. **Undercoverage bias** occurs when some individuals of the population are inherently less likely to be included in the sample than others, making the sample not representative of the population. In the example, this bias creates a problem because a degree in health-related fields might take more or less time to complete than a degree in other fields. Suppose that it takes longer. Since graduates from other fields would be less likely to be in the sample, the undercoverage bias would cause her to *overestimate* the parameter.

Sampling randomly resolves the problem of undercoverage bias, *if the sample is randomly selected from the entire population of interest*. If the sample is randomly selected from only a subset of the population, say, only graduates from health-related fields, then the sample will not be representative of the population of interest. Generalizations can only be made to the population from which the sample is randomly selected.

The most basic random sample is called a **simple random sample**, which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

A common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.
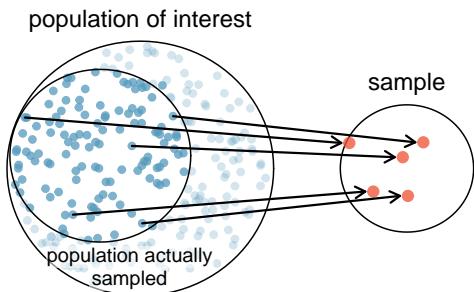
Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Similarly, a **volunteer sample** is one in which people's responses are solicited and those who choose to participate, respond. This is a problem because those who choose to participate may tend to have different opinions than the rest of the population, resulting in a biased sample.

GUIDED PRACTICE 1.20

We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[28]

The act of taking a random sample helps minimize bias; however, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.

Even if a sample has no undercoverage bias and no non-response bias, there is an additional type of bias that often crops up and undermines the validity of results, known as response bias. **Response bias** refers to a broad range of factors that influence how a person responds, such as question wording, question order, and influence of the interviewer. This type of bias can be present even when we collect data from an entire population in what is called a **census**. Because response bias is often subtle, one must pay careful attention to how questions were asked when attempting to draw conclusions from the data.

EXAMPLE 1.21

Suppose a high school student wants to investigate the student body's opinions on the food in the cafeteria. Let's assume that she manages to survey every student in the school. How might response bias arise in this context?

———————

There are many possible correct answers to this question. For example, students might respond differently depending upon who asks the question, such as a school friend or someone who works in the cafeteria. The wording of the question could introduce response bias. Students would likely respond differently if asked "Do you like the food in the cafeteria?" versus "The food in the cafeteria is pretty bad, don't you think?"

WATCH OUT FOR BIAS

Undercoverage bias, non-response bias, and response bias can still exist within a random sample. Always determine how a sample was chosen, ask what proportion of people failed to respond, and critically examine the wording of the questions.

When there is no bias in a sample, increasing the sample size tends to increase the precision and reliability of the estimate. When a sample is biased, it may be impossible to decipher helpful information from the data, even if the sample is very large.

---

[28]Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind.

GUIDED PRACTICE 1.22

A researcher sends out questionnaires to 50 randomly selected households in a particular town asking whether or not they support the addition of a traffic light in their neighborhood. Because only 20% of the questionnaires are returned, she decides to mail questionnaires to 50 more randomly selected households in the same neighborhood. Comment on the usefulness of this approach.[29]

1.4.3   Simple, systematic, stratified, cluster, and multistage sampling

Almost all statistical methods for observational data rely on a sample being random and unbiased. When a sample is collected in a biased way, these statistical methods will not generally produce reliable information about the population.

The idea of a simple random sample was introduced in the last section. Here we provide a more technical treatment of this method and introduce four new random sampling methods: systematic, stratified, cluster, and multistage.[30] Figure 1.14 provides a graphical representation of simple versus systematic sampling while Figure 1.15 provides a graphical representation of stratified, cluster, and multistage sampling.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. For the 2019 season, N, the population size or total number of players, is 750. To take a simple random sample of $n = 120$ of these baseball players and their salaries, we could number each player from 1 to 750. Then we could randomly select 120 numbers between 1 and 750 (without replacement) using a random number generator or random digit table. The players with the selected numbers would comprise our sample.

Two properties are always true in a simple random sample:

1. Each case in the population has an equal chance of being included in the sample.

2. Each *group* of n cases has an equal chance of making up the sample.

The statistical methods in this book focus on data collected using simple random sampling. Note that Property 2 – that each group of n cases has an equal chance making up the sample – is not true for the remaining four sampling techniques. As you read each one, consider why.

Though less common than simple random sampling, **systematic sampling** is sometimes used when there exists a convenient list of all of the individuals of the population. Suppose we have a roster with the names of all the MLB players from the 2019 season. To take a systematic random sample, number them from 1 to 750. Select one random number between 1 and 750 and let that player be the first individual in the sample. Then, depending on the desired sample size, select every 10th number or 20th number, for example, to arrive at the sample.[31] If there are no patterns in the salaries based on the numbering then this could be a reasonable method.

EXAMPLE 1.23

A systematic sample is not the same as a simple random sample. Provide an example of a sample that can come from a simple random sample but not from a systematic random sample.

———

Answers can vary. If we take a sample of size 3, then it is possible that we could sample players numbered 1, 2, and 3 in a simple random sample. Such a sample would be impossible from a systematic sample. Property 2 of simple random samples does not hold for other types of random samples.

---

[29]The researcher should be concerned about non-response bias, and sampling more people will not eliminate this issue. The same type of people that did not respond to the first survey are likely not going to respond to the second survey. Instead, she should make an effort to reach out to the households from the original sample that did not respond and solicit their feedback, possibly by going door-to-door.

[30]Multistage sampling is not part of the AP syllabus.

[31]If we want a sample of size $n = 150$, it would make sense to select every 5th player since $750/150 = 5$. Suppose we randomly select the number 741. Then player 741, 746, 1, 6, 11, $\cdots$, 731, and 736 would make up the sample.
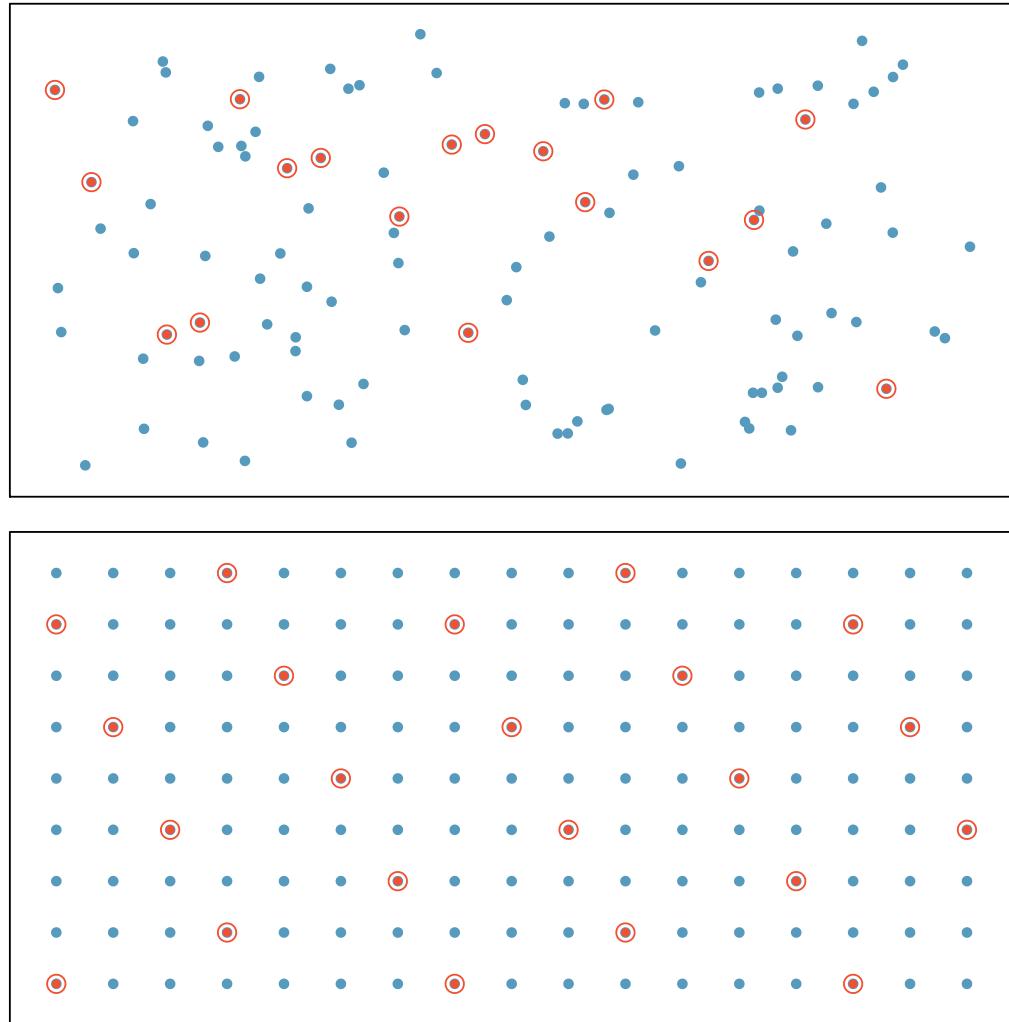
Figure 1.14: Examples of simple random sampling and systematic sampling. In the top panel, simple random sampling was used to randomly select 18 cases. In the lower panel, systematic random sampling was used to select every 7th individual.

Sometimes there is a variable that is known to be associated with the quantity we want to estimate. In this case, a stratified random sample might be selected. **Stratified sampling** is a divide-and-conquer sampling strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together and a sampling method, usually simple random sampling, is employed to select a certain number or a certain proportion of the whole within each stratum. In the baseball salary example, the 30 teams could represent the strata; some teams have a lot more money (we're looking at you, Yankees).

EXAMPLE 1.24

For this baseball example, briefly explain how to select a stratified random sample of size $n = 120$.
————————

Each team can serve as a stratum, and we could take a simple random sample of 4 players from each of the 30 teams, yielding a sample of 120 players.

Stratified sampling is inherently different than simple random sampling. For example, the stratified sampling approach described would make it impossible for the entire Yankees team to be included in the sample.

EXAMPLE 1.25

Stratified sampling is especially useful when the cases in each stratum are very similar *with respect to the outcome of interest*. Why is it good for cases within each stratum to be very similar?
————————

We should get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population. For example, in a simple random sample, it is possible that just by random chance we could end up with proportionally too many Yankees players in our sample, thus overestimating the true average salary of all MLB players. A stratified random sample can assure proportional representation from each team.

Next, let's consider a sampling technique that randomly selects groups of people. **Cluster sampling** is much like simple random sampling, but instead of randomly selecting *individuals*, we randomly select groups or **clusters**. Unlike stratified sampling, cluster sampling is most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. That is, we expect individual strata to be **homogeneous** (self-similar), while we expect individual clusters to be **heterogeneous** (diverse) with respect to the variable of interest.

Sometimes cluster sampling can be a more economical random sampling technique than the alternatives. For example, if neighborhoods represented clusters, this sampling method works best when each neighborhood is very diverse. Because each neighborhood itself encompasses diversity, a cluster sample can reduce the time and cost associated with data collection, because the interviewer would need only go to some of the neighborhoods rather than to all parts of a city, in order to collect a useful sample.

Multistage sampling, also called **multistage cluster sampling**, is a two (or more) step strategy. The first step is to take a cluster sample, as described above. Then, instead of including all of the individuals in these clusters in our sample, a second sampling method, usually simple random sampling, is employed within each of the selected clusters. In the neighborhood example, we could first randomly select some number of neighborhoods and then take a simple random sample from just those selected neighborhoods. As seen in Figure 1.15, stratified sampling requires observations to be sampled from *every* stratum. Multistage sampling selects observations *only* from those clusters that were randomly selected in the first step.

It is also possible to have more than two steps in multistage sampling. Each cluster may be naturally divided into subclusters. For example, each neighborhood could be divided into streets. To take a three-stage sample, we could first select some number of clusters (neighborhoods), and then, within the selected clusters, select some number of subclusters (streets). Finally, we could select some number of individuals from each of the selected streets.
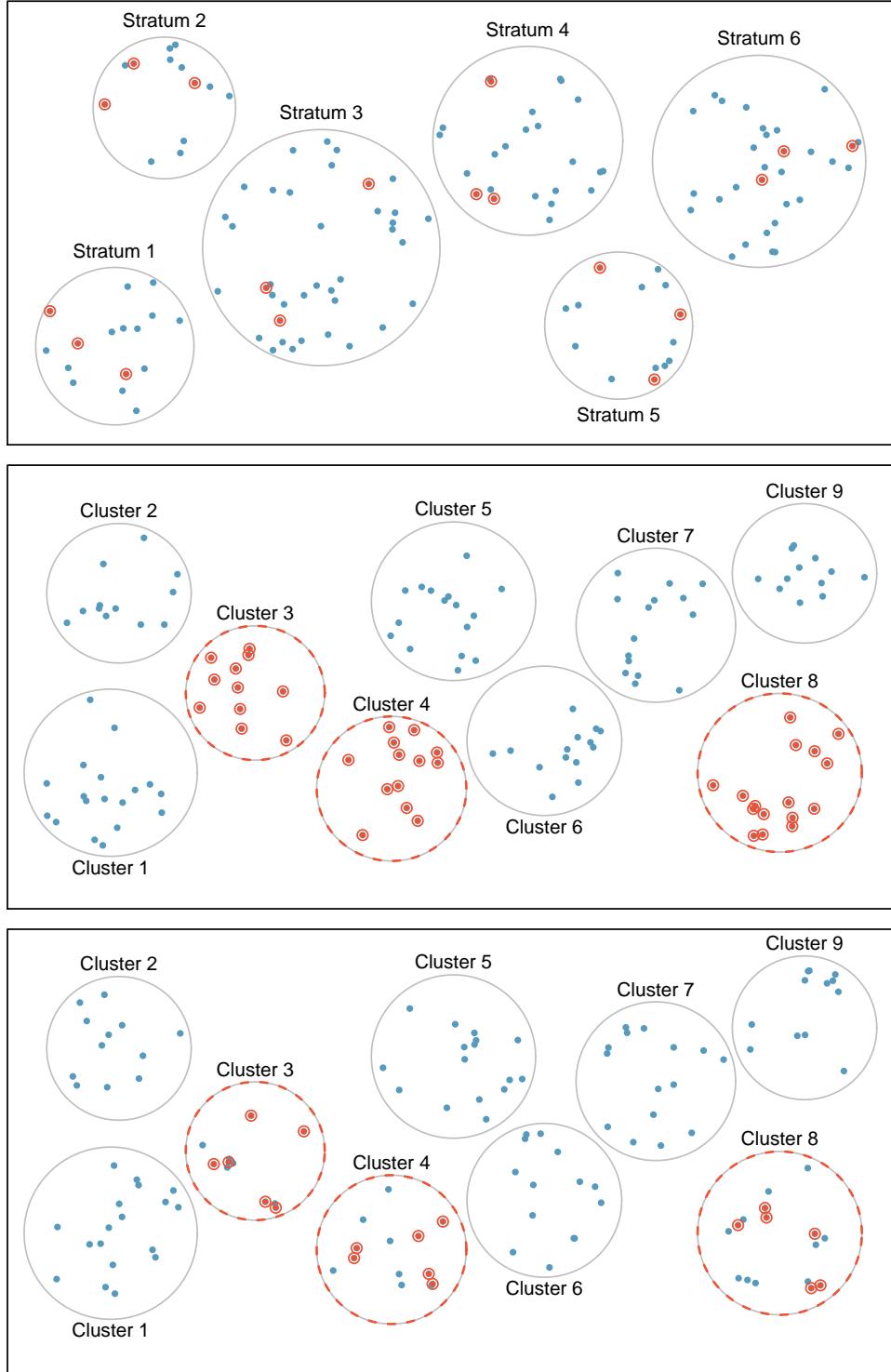
Figure 1.15: Examples of stratified, cluster, and multistage sampling. In the top panel, stratified sampling was used: cases were grouped into strata, and then simple random sampling was employed within each stratum. In the middle panel, cluster sampling was used, where data were binned into nine cluster and three clusters were randomly selected. In the bottom panel, multistage sampling was used. Data were binned into the nine clusters, three of the cluster were randomly selected, and then six cases were randomly sampled in each of the three selected clusters.

EXAMPLE 1.26

Suppose we are interested in estimating the proportion of students at a certain school that have part-time jobs. It is believed that older students are more likely to work than younger students. What sampling method should be employed? Describe how to collect such a sample to get a sample size of 60.

───────────

Because grade level affects the likelihood of having a part-time job, we should take a stratified random sample. To do this, we can take a simple random sample of 15 students from each grade. This will give us equal representation from each grade. Note: in a simple random sample, just by random chance we might get too many students who are older or younger, which could make the estimate too high or too low. Also, there are no well-defined clusters in this example. We wouldn't want to use the grades as clusters and sample everyone from a couple of the grades. This would create too large a sample and would not give us the nice representation from each grade afforded by the stratified random sample.

EXAMPLE 1.27

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

───────────

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, multistage cluster sampling seems like a very good idea. First, we might randomly select half the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample and would still give us reliable information.

ADVANCED SAMPLING TECHNIQUES REQUIRE ADVANCED METHODS

The methods of inference covered in this book generally only apply to simple random samples. More advanced analysis techniques are required for systematic, stratified, cluster, and multistage random sampling.

Section summary

- In an **observational study**, one must always consider the existence of **confounding factors**. A confounding factor is a "spoiler variable" that could explain an observed relationship between the explanatory variable and the response. Remember: For a variable to be confounding it must be associated with both the explanatory variable *and* the response variable.

- When taking a sample from a population, avoid **convenience samples** and **volunteer samples**, which likely introduce bias. Instead, use a **random** sampling method.

- Generalizations from a sample can be made to a population only if the sample is random. Furthermore, the generalization can be made only to the population from which the sample was randomly selected, not to a larger or different population.

- Random sampling from the entire population of interest avoids the problem of **undercoverage bias**. However, **response bias** and **non-response** bias can be present in any type of sample, random or not.

- In a **simple random sample**, every *individual* as well as every *group of individuals* has the same probability of being in the sample. A common way to select a simple random sample is to number each individual of the population from 1 to N. Using a random digit table or a random number generator, numbers are randomly selected without replacement and the corresponding individuals become part of the sample.

- A **systematic random sample** involves choosing from of a population using a random starting point, and then selecting members according to a fixed, periodic interval (such as every 10th member).

- A **stratified random sample** involves randomly sampling from *every* **strata**, where the strata should correspond to a variable thought to be associated with the variable of interest. This ensures that the sample will have appropriate representation from each of the different strata and reduces variability in the sample estimates.

- A **cluster random sample** involves randomly selecting a set of **clusters**, or groups, and then collecting data on all individuals in the selected clusters. This can be useful when sampling clusters is more convenient and less expensive than sampling individuals, and it is an effective strategy when each cluster is approximately representative of the population.

- Remember: *Individual strata should be homogeneous (self-similar), while individual clusters should be heterogeneous (diverse).* For example, if smoking is correlated with what is being estimated, let one stratum be all smokers and the other be all non-smokers, then randomly select an appropriate number of *individuals* from *each* strata. Alternately, if age is correlated with the variable being estimated, one could randomly select a *subset* of clusters, where each cluster has mixed age groups.

Exercises

1.19   Course satisfaction across sections. A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.
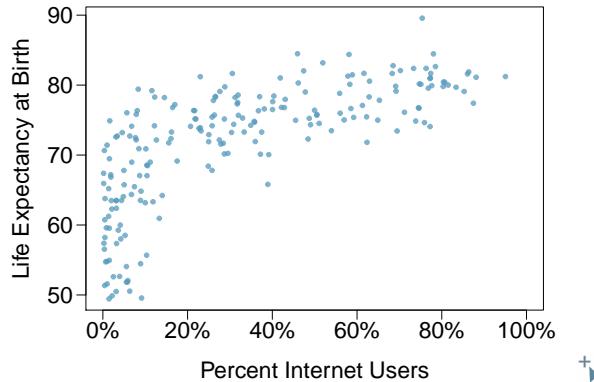
(a) What type of study is this?

(b) Suggest a sampling strategy for carrying out this study.

1.20   Housing proposal across dorms. On a large college campus first-year students and sophomores live in dorms located on the eastern part of the campus and juniors and seniors live in dorms located on the western part of the campus. Suppose you want to collect student opinions on a new housing structure the college administration is proposing and you want to make sure your survey equally represents opinions from students from all years.

(a) What type of study is this?

(b) Suggest a sampling strategy for carrying out this study.

1.21   Internet use and life expectancy. The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.[32]

(a) Describe the relationship between life expectancy and percentage of internet users.

(b) What type of study is this?

(c) State a possible confounding variable that might explain this relationship and describe its potential effect.

1.22   Stressed out, Part I. A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

(c) State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

1.23   Evaluate sampling methods. A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not.

(a) Survey a simple random sample of 500 students.

(b) Stratify students by their field of study, then sample 10% of students from each stratum.

(c) Cluster students by their ages (e.g. 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

1.24   Random digit dialing. The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

---

[32]CIA Factbook, Country Comparisons, 2014.

1.25 Haters are gonna hate, study confirms. A study published in the *Journal of Personality and Social Psychology* asked a group of 200 randomly sampled men and women to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their dispositional attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively also tended to react negatively to it. Researchers concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."[33]

(a) What are the cases?

(b) What is (are) the response variable(s) in this study?

(c) What is (are) the explanatory variable(s) in this study?

(d) Does the study employ random sampling?

(e) Is this an observational study or an experiment? Explain your reasoning.

(f) Can we establish a causal link between the explanatory and response variables?

(g) Can the results of the study be generalized to the population at large?

1.26 Family size. Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

1.27 Sampling strategies. A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

(a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.

(b) He gives out the survey only to his friends, making sure each one of them fills out the survey.

(c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey.

(d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

1.28 Reading the paper. Below are excerpts from two articles published in the *NY Times*:

(a) An article titled *Risks: Smokers Found More Prone to Dementia* states the following:[34]

> "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

(b) Another article titled *The School Bully Is Sleepy* states the following:[35]

> "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

---

[33]Justin Hepler and Dolores Albarracín. "Attitudes without objects - Evidence for a dispositional attitude, its measurement, and its consequences". In: *Journal of personality and social psychology* 104.6 (2013), p. 1060.

[34]R.C. Rabin. "Risks: Smokers Found More Prone to Dementia". In: *New York Times* (2010).

[35]T. Parker-Pope. "The School Bully Is Sleepy". In: *New York Times* (2011).

1.5   Experiments

You would like to determine if drinking a cup of tea each morning will cause students to perform better on tests. What are different ways you could design an experiment to answer this question? What are possible sources of bias, and how would you try to minimize them? The goal of an experiment is to be able to draw a causal conclusion about the effect of a treatment – in this case, drinking tea. If the design is poor, a causal conclusion cannot be drawn, even if you observe an association between drinking tea and performing better on tests. This is why it is crucial to start with a well-designed experiment.

Learning objectives

1. Identify the subjects/experimental units, treatments, and response variable in an experiment.

2. Identify the three main principles of experiment design and explain their purpose: direct control, randomization, and replication.

3. Explain placebo effect and describe when and how to implement a single-blind and a double-blind experiment.

4. Identify and describe how to implement the following three experimental designs: completely randomized design, blocked design, and matched pairs design.

5. Explain the purpose of random assignment or randomization in each of the three experimental designs.

6. Explain how to randomize treatments in a completely randomized design using technology or a table of random digits (make sure this is explained).

7. Explain when it is reasonable to draw a causal conclusion about the effect of a treatment.

8. Identify the number of factors in experiment, the number of levels for each factor and the total number of treatments.

1.5.1   Reducing bias in human experiments

In the last section we investigated observational studies and sampling strategies. While these are effective tools for answering certain research questions, often times researchers want to measure the effect of a treatment. In this case, they must carry out an experiment. Just as randomization is essential in sampling in order to avoid selection bias, randomization is essential in the context of experiments to determine which subjects will receive which treatments. If the researcher chooses which patients are in the treatment and control groups, she may unintentionally place sicker patients in the treatment group, biasing the experiment against the treatment.

Randomized experiments are essential for investigating cause and effect relationships, but they do not ensure an unbiased perspective in all cases. Human studies are perfect examples where bias

can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients.[36] In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[37] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment. In an experiment, the explanatory variable is also called a **factor**. Here the factor is receiving the drug treatment. It has two **levels**: yes and no, thus it is categorical. The response variable is whether or not patients died within the time frame of the study. It is also categorical.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be **blind** or **single-blind**. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a **placebo**, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the **placebo effect**.

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where researchers who interact with subjects and are responsible for measuring the response variable are, just like the subjects, unaware of who is or is not receiving the treatment.[38]

GUIDED PRACTICE 1.28

Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[39]

1.5.2 Principles of experimental design

Well-conducted experiments are built on three main principles.

Direct Control. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. They want the groups to be as identical as possible *except for the treatment*, so that at the end of the experiment any difference in response between the groups can be attributed to the treatment and not to some other confounding or lurking variable. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

---

[36]Anturane Reinfarction Trial Research Group. 1980. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.

[37]Human subjects are often called **patients**, **volunteers**, or **study participants**.

[38]There are always some researchers involved in the study who do know which patients are receiving which treatment. However, they do not interact with the study's patients and do not tell the blinded health care professionals who is receiving which treatment.

[39]The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients could distinguish what treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

Direct control refers to variables that the researcher can control, or make the same. A researcher can directly control the appearance of the treatment, the time of day it is taken, etc. She cannot directly control variables such as gender or age. To control for these other types of variables, she might consider blocking, which is described in Section 1.5.3.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps *even out* the effects of such differences, and it also prevents accidental bias from entering the study.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In an experiment with six subjects, even if there is randomization, it is quite possible for the three healthiest people to be in the same treatment group. In a randomized experiment with 100 people, it is virtually impossible for the healthiest 50 people to end up in the same treatment group. In a single study, we **replicate** by imposing the treatment on a sufficiently large number of subjects or experimental units. A group of scientists may also replicate an entire study to verify an earlier finding. However, each study should ensure a sufficiently large number of subjects because, in many cases, there is no opportunity or funding to carry out the entire experiment again.

It is important to incorporate these design principles into any experiment. If they are lacking, the inference methods presented in the following chapters will not be applicable and their results may not be trustworthy. In the next section we will consider three types of experimental design.

1.5.3   Completely randomized, blocked, and matched pairs design

A **completely randomized experiment** is one in which the subjects or experimental units are randomly assigned to each group in the experiment. Suppose we have three treatments, one of which may be a placebo, and 300 subjects. To carry out a completely randomized design, we could randomly assign each subject a unique number from 1 to 300, then subjects with numbers 1-100 would get treatment 1, subjects 101-200 would get treatment 2, and subjects 201- 300 would get treatment 3. Note that this method of randomly allocating subjects to treatments in not equivalent to taking a simple random sample. Here we are not sampling a subset of a population; we are randomly *splitting* subjects into groups.

While it might be ideal for the subjects to be a random sample of the population of interest, that is rarely the case. Subjects must volunteer to be part of an experiment. However, because randomization is incorporated in the splitting of the groups, we can still use statistical techniques to check for a causal connection, though the precise population for which the conclusion applies may be unclear. For example, if an experiment to determine the most effective means to encourage individuals to vote is carried out only on college students, we may not be able to generalize the conclusions of the experiment to all adults in the population.

Researchers sometimes know or suspect that another variable, other than the treatment, influences the response. Under these circumstances, they may carry out a **blocked experiment**. In this design, they first group individuals into **blocks** based on the identified variable and then randomize subjects within each block to the treatment groups. This strategy is referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks. Then we can randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.16. At the end of the experiment, we would incorporate this blocking into the analysis. By blocking by risk of patient, we control for this possible confounding factor. Additionally, by randomizing subjects to treatments within each block, we attempt to even out the effect of variables that we cannot block or directly control.
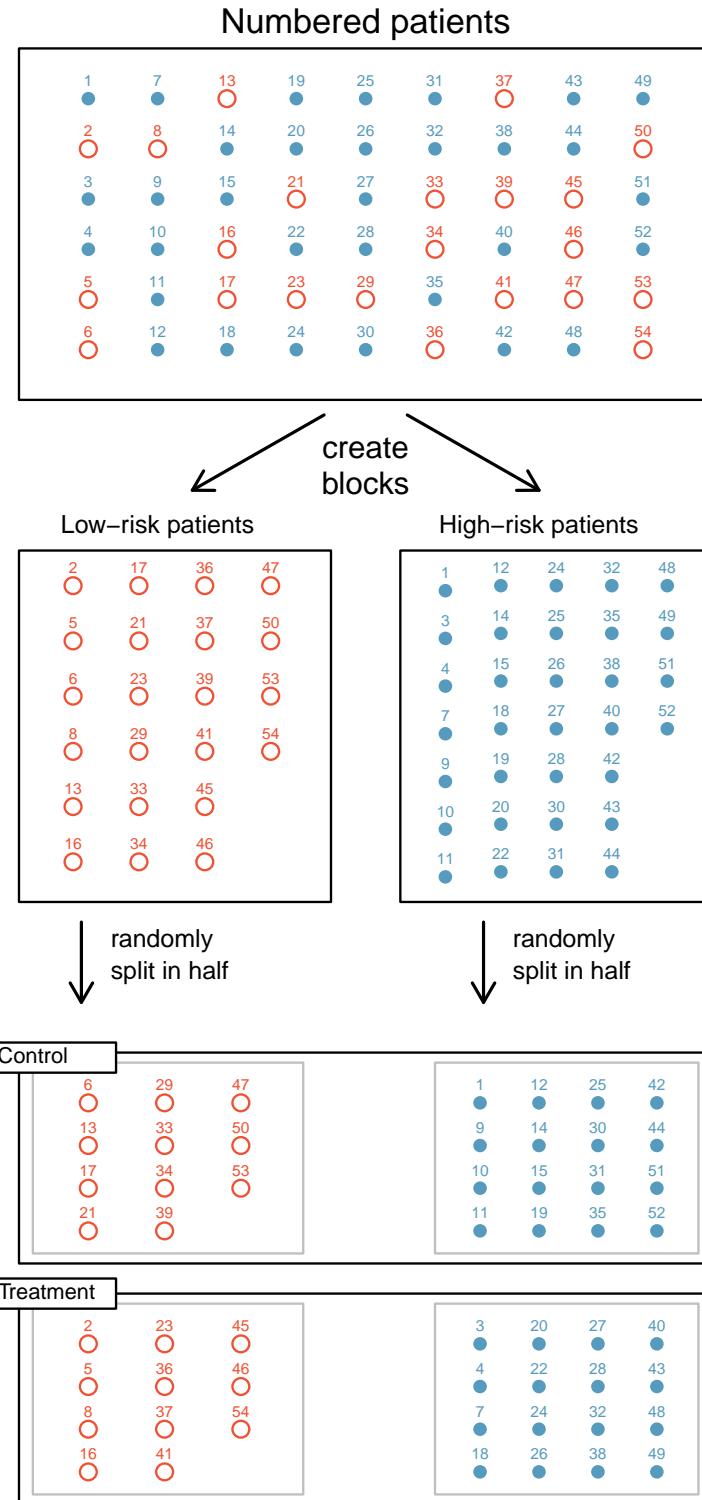
Figure 1.16: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

EXAMPLE 1.29

An experiment will be conducted to compare the effectiveness of two methods for quitting smoking. Identify a variable that the researcher might wish to use for blocking and describe how she would carry out a blocked experiment.

––––––––––

The researcher should choose the variable that is most likely to influence the response variable - whether or not a smoker will quit. A reasonable variable, therefore, would be the number of years that the smoker has been smoking. The subjects could be separated into three blocks based on number of years of smoking and each block randomly divided into the two treatment groups.

Even in a blocked experiment with randomization, other variables that affect the response can be distributed unevenly among the treatment groups, thus biasing the experiment in one direction. A third type of design, known as **matched pairs** addresses this problem. In a matched pairs experiment, pairs of people are matched on as many variables as possible, so that the comparison happens between very similar cases. This is actually a special type of blocked experiment, where the blocks are of size two.

An alternate form of matched pairs involves each subject receiving *both* treatments. Randomization can be incorporated by randomly selecting half the subjects to receive treatment 1 first, followed by treatment 2, while the other half receives treatment 2 first, followed by treatment.

GUIDED PRACTICE 1.30

How and why should randomization be incorporated into a matched pairs design?[40]

GUIDED PRACTICE 1.31

Matched pairs sometimes involves each subject receiving both treatments at the same time. For example, if a hand lotion was being tested, half of the subjects could be randomly assigned to put Lotion A on the left hand and Lotion B on the right hand, while the other half of the subjects would put Lotion B on the left hand and Lotion A on the right hand. Why would this be a better design than a completely randomized experiment in which half of the subjects put Lotion A on both hands and the other half put Lotion B on both hands?[41]

Because it is essential to identify the type of data collection method used when choosing an appropriate inference procedure, we will revisit sampling techniques and experiment design in the subsequent chapters on inference.

––––––––––––––––––––––––––––

[40]Assume that all subjects received treatment 1 first, followed by treatment 2. If the variable being measured happens to increase naturally over the course of time, it would appear as though treatment 2 had a greater effect than it really did.

[41]The dryness of people's skins varies from person to person, but probably less so from one person's right hand to left hand. With the matched pairs design, we are able control for this variability by comparing each person's right hand to her left hand, rather than comparing some people's hands to other people's hands (as you would in a completely randomized experiment).

1.5.4 Testing more than one variable at a time

Some experiments study more than one factor (explanatory variable) at a time, and each of these factors may have two or more levels (possible values). For example, suppose a researcher plans to investigate how the type and volume of music affect a person's performance on a particular video game. Because these two factors, `type` and `volume`, could interact in interesting ways, we do not want to test one factor at a time. Instead, we want to do an experiment in which we test all the *combinations* of the factors. Let's say that `volume` has two levels (soft and loud) and that `type` has three levels (dance, classical, and punk). Then, we would want to carry out the experiment at each of the six (2 x 3 = 6) combinations: soft dance, soft classical, soft punk, loud dance, loud classical, loud punk. Each of the these combinations is a **treatment**. Therefore, this experiment will have 2 factors and 6 treatments. In order to replicate each treatment 10 times, one would need to play the game 60 times.

GUIDED PRACTICE 1.32

A researcher wants to compare the effectiveness of four different drugs. She also wants to test each of the drugs at two doses: low and high. Describe the factors, levels, and treatments of this experiment.[42]

As the number of factors and levels increases, the number of treatments become large and the analysis of the resulting data becomes more complex, requiring the use of advanced statistical methods. We will investigate only one factor at a time in this book.

---

[42]There are two factors: type of drug, which has four levels, and dose, which has 2 levels. There will be 4 x 2 = 8 treatments: drug 1 at low dose, drug 1 at high dose, drug 2 at low dose, and so on.

Section summary

- In an **experiment**, researchers impose a **treatment** to test its effects. In order for observed differences in the response to be attributed to the treatment and not to some other factor, it is important to make the treatment groups and the conditions for the treatment groups as similar as possible.

- Researchers use **direct control**, ensuring that variables that are within their power to modify (such as drug dosage or testing conditions) are made the *same* for each treatment group.

- Researchers **randomly** assign subjects to the treatment groups so that the effects of uncontrolled and potentially confounding variables are *evened out* among the treatment groups.

- **Replication**, or imposing the treatments on many subjects, gives more data and decreases the likelihood that the treatment groups differ on some characteristic due to chance alone (i.e. in spite of the randomization).

- An ideal experiment is **randomized**, **controlled**, and **double-blind**.

- A **completely randomized experiment** involves randomly assigning the subjects to the different treatment groups. To do this, first number the subjects from 1 to N. Then, randomly choose some of those numbers and assign the corresponding subjects to a treatment group. Do this in such a way that the treatment group sizes are balanced, unless there exists a good reason to make one treatment group larger than another.

- In a **blocked experiment**, subjects are first separated by a variable thought to affect the response variable. Then, within *each* block, subjects are randomly assigned to the treatment groups as described above, allowing the researcher to compare like to like within each block.

- When feasible, a **matched-pairs experiment** is ideal, because it allows for the best comparison of like to like. A matched-pairs experiment can be carried out on pairs of subjects that are meaningfully paired, such as twins, or it can involve all subjects receiving both treatments, allowing subjects to be compared to *themselves*.

- A treatment is also called a **factor** or explanatory variable. Each treatment/factor can have multiple **levels**, such as yes/no or low/medium/high. When an experiment includes many factors, multiplying the number of levels of the factors together gives the total number of treatment groups.

- In an experiment, blocking, randomization, and direct control are used to *control for confounding factors*.

Exercises

1.29 Light and exam performance. A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

(a) What is the response variable?
(b) What is the explanatory variable? What are its levels?
(c) What is the blocking variable? What are its levels?

1.30 Vitamin supplements. To assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.[43]

(a) Was this an experiment or an observational study? Why?
(b) What are the explanatory and response variables in this study?
(c) Were the patients blinded to their treatment?
(d) Was this study double-blind?
(e) Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

1.31 Light, noise, and exam performance. A study is designed to test the effect of light level and noise level on exam performance of students. The researcher believes that light and noise levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The light treatments considered are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). The noise treatments considered are no noise, construction noise, and human chatter noise.

(a) What type of study is this?
(b) How many factors are considered in this study? Identify them, and describe their levels.
(c) What is the role of the sex variable in this study?

1.32 Music and learning. You would like to conduct an experiment in class to see if students learn better if they study without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

1.33 Soda preference. You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.

1.34 Exercise and mental health. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?
(b) What are the treatment and control groups in this study?
(c) Does this study make use of blocking? If so, what is the blocking variable?
(d) Does this study make use of blinding?
(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.
(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

---

[43]C. Audera et al. "Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial". In: *Medical Journal of Australia* 175.7 (2001), pp. 359–362.

Chapter highlights

Chapter 1 focused on various ways that researchers collect data. The key concepts are the difference between a sample and an experiment and the role that randomization plays in each.

- Researchers take a **random sample** in order to draw an **inference** to the larger population from which they sampled. When examining observational data, even if the individuals were randomly sampled, a correlation does not imply a causal link.

- In an **experiment**, researchers impose a treatment and use **random assignment** in order to draw **causal conclusions** about the effects of the treatment. While often implied, inferences to a larger population may not be valid if the subjects were not also *randomly sampled* from that population.

Related to this are some important distinctions regarding terminology. The terms stratifying and blocking cannot be used interchangeably. Likewise, taking a simple random sample is different than randomly assigning individuals to treatment groups.
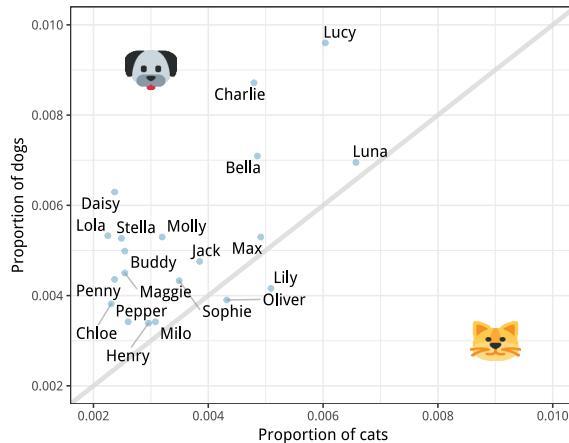
- **Stratifying vs Blocking**. Stratifying is used when sampling, where the purpose is to *sample* a subgroup from each stratum in order to arrive at a better *estimate* for the parameter of interest. Blocking is used in an experiment to *separate* subjects into blocks and then *compare* responses within those blocks. All subjects in a block are used in the experiment, not just a sample of them.

- **Random sampling vs Random assignment**. Random sampling refers to sampling a subset of a population for the purpose of inference to that population. Random assignment is used in an experiment to separate subjects into groups for the purpose of comparison between those groups.

When randomization is not employed, as in an **observational study**, neither inferences nor causal conclusions can be drawn. Always be mindful of possible **confounding factors** when interpreting the results of observation studies.

Chapter exercises

1.35 Pet names. The city of Seattle, WA has an open data portal that includes pets registered in the city. For each registered pet, we have information on the pet's name and species. The following visualization plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the $x = y$ line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.

(a) Are these data collected as part of an experiment or an observational study?

(b) What is the most common dog name? What is the most common cat name?

(c) What names are more common for cats than dogs?

(d) Is the relationship between the two variables positive or negative? What does this mean in context of the data?

1.36 Stressed out, Part II. In a study evaluating the relationship between stress and muscle cramps, half the subjects are randomly assigned to be exposed to increased stress by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

1.37 Chia seeds and weight loss. Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.[44]

(a) What type of study is this?

(b) What are the experimental and control treatments in this study?

(c) Has blocking been used in this study? If so, what is the blocking variable?

(d) Has blinding been used in this study?

(e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

1.38 City council survey. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context.

(a) Randomly sample 200 households from the city.

(b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.

(c) Divide the city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.

(d) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

(e) Sample the 200 households closest to the city council offices.

---
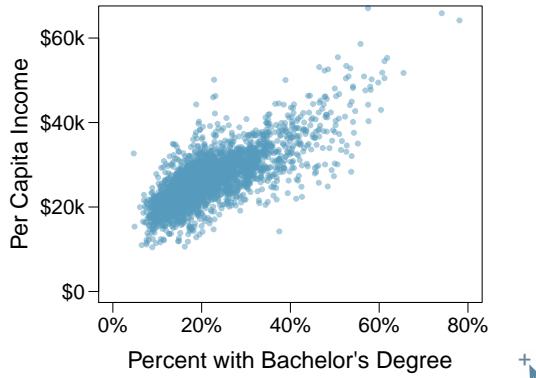
[44]D.C. Nieman et al. "Chia seed does not promote weight loss or alter disease risk factors in overweight adults". In: *Nutrition Research* 29.6 (2009), pp. 414–418.

1.39   Flawed reasoning. Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

(a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

(b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

(c) An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

1.40   Income and education in US counties. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,142 counties in the US using American Community Survey data from 2017.

(a) What are the explanatory and response variables?

(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

(c) Can we conclude that having a bachelor's degree increases one's income?

1.41   Eat better, feel better? In a public health study on the effects of consumption of fruits and vegetables on psychological well-being in young adults, participants were randomly assigned to three groups: (1) diet-as-usual, (2) an ecological momentary intervention involving text message reminders to increase their fruits and vegetable consumption plus a voucher to purchase them, or (3) a fruit and vegetable intervention in which participants were given two additional daily servings of fresh fruits and vegetables to consume on top of their normal diet. Participants were asked to take a nightly survey on their smartphones. Participants were student volunteers at the University of Otago, New Zealand. At the end of the 14-day study, only participants in the third group showed improvements to their psychological well-being across the 14-days relative to the other groups.[45]

(a) What type of study is this?

(b) Identify the explanatory and response variables.

(c) Comment on whether the results of the study can be generalized to the population.

(d) Comment on whether the results of the study can be used to establish causal relationships.

(e) A newspaper article reporting on the study states, "The results of this study provide proof that giving young adults fresh fruits and vegetables to eat can have psychological benefits, even over a brief period of time." How would you suggest revising this statement so that it can be supported by the study?

---

[45]Tamlin S Conner et al. "Let them eat fruit! The effect of fruit and vegetable consumption on psychological well-being in young adults: A randomized controlled trial". In: *PloS one* 12.2 (2017), e0171206.

1.42   Screens, teens, and psychological well-being. In a study of three nationally representative large-scale data sets from Ireland, the United States, and the United Kingdom (n = 17,247), teenagers between the ages of 12 to 15 were asked to keep a diary of their screen time and answer questions about how they felt or acted. The answers to these questions were then used to compute a psychological well-being score. Additional data were collected and included in the analysis, such as each child's sex and age, and on the mother's education, ethnicity, psychological distress, and employment. The study concluded that there is little clear-cut evidence that screen time decreases adolescent well-being.[46]

(a) What type of study is this?
(b) Identify the explanatory variables.
(c) Identify the response variable.
(d) Comment on whether the results of the study can be generalized to the population, and why.
(e) Comment on whether the results of the study can be used to establish causal relationships.

1.43   Stanford Open Policing. The Stanford Open Policing project gathers, analyzes, and releases records from traffic stops by law enforcement agencies across the United States. Their goal is to help researchers, journalists, and policymakers investigate and improve interactions between police and the public.[47] The following is an excerpt from a summary table created based off of the data collected as part of this project.

County	State	Driver's race	No. of stops per year	% of stopped cars searched	drivers arrested
Apaice County	Arizona	Black	266	0.08	0.02
Apaice County	Arizona	Hispanic	1008	0.05	0.02
Apaice County	Arizona	White	6322	0.02	0.01
Cochise County	Arizona	Black	1169	0.05	0.01
Cochise County	Arizona	Hispanic	9453	0.04	0.01
Cochise County	Arizona	White	10826	0.02	0.01
...	...	...	...	...	...
Wood County	Wisconsin	Black	16	0.24	0.10
Wood County	Wisconsin	Hispanic	27	0.04	0.03
Wood County	Wisconsin	White	1157	0.03	0.03

(a) What variables were collected on each individual traffic stop in order to create to the summary table above?
(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.
(c) Suppose we wanted to evaluate whether vehicle search rates are different for drivers of different races. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

1.44   Space launches. The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).[48]

	1957 - 1999		2000 - 2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	–	–	5	65

(a) What variables were collected on each launch in order to create to the summary table above?
(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.
(c) Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

---

[46]Amy Orben and AK Baukney-Przybylski. "Screens, Teens and Psychological Well-Being: Evidence from three time-use diary studies". In: *Psychological Science* (2018).

[47]Emma Pierson et al. "A large-scale analysis of racial disparities in police stops across the United States". In: *arXiv preprint arXiv:1706.05678* (2017).

[48]JSR Launch Vehicle Database, A comprehensive list of suborbital space launches, 2019 Feb 10 Edition.

Chapter 2

Summarizing data

After collecting data, the next stage in the investigative process is to describe and summarize the data. In this chapter, we will look at ways to summarize numerical and categorical data graphically, numerically, and verbally. While in practice, numerical and graphical summaries are done using computer software, it is helpful to understand how these summaries are created and it is especially important to understand how to interpret and communicate these findings.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

2.1   Examining numerical data

How do we visualize and describe the distribution of household income for counties within the United States? What shape would the distribution have? What other features might be important to notice? In this section, we will explore techniques for summarizing numerical variables. We will apply these techniques using county-level data from the US Census Bureau, which was introduced in Section 1.2, and a new data set `email50`, that comprises information on a random sample of 50 emails.

Learning objectives

1. Use scatterplots to represent bivariate data and to see the relationship between two numerical variables. Describe the direction, form, and strength of the relationship, as well as any unusual observations.

2. Understand what the term distribution means and how to summarize it in a table or a graph.

3. Create univariate displays, including stem-and-leaf plots, dot plots, and histograms, to visualize the distribution of a numerical variable. Be able to read off specific information and summary information from these graphs.

4. Identify the shape of a distribution as approximately symmetric, right skewed, or left skewed. Also, identify whether a distribution is unimodal, bimodal, multimodal, or uniform.

5. Read and interpret a cumulative frequency or cumulative relative frequency histogram.

2.1.1   Scatterplots for paired data

Sometimes researchers wish to see the relationship between two variables. When we talk of a relationship or an association between variables, we are interested in how one variable behaves as the other variable increases or decreases.

A **scatterplot** provides a case-by-case view of data that illustrates the relationship between two numerical variables. A scatterplot is shown in Figure 2.1, illustrating the relationship between the number of line breaks (`line_breaks`) and number of characters (`num_char`) in emails for the `email50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figure 2.1.
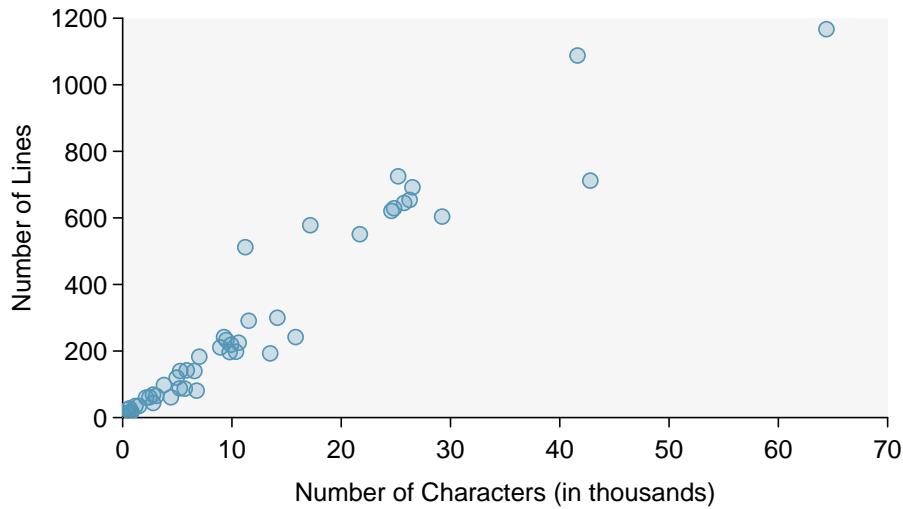
Figure 2.1: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

EXAMPLE 2.1

A scatterplot requires **bivariate**, or **paired data**. What does paired data mean?

We say observations are *paired* when the two observations correspond to the same case or individual. In unpaired data, there is no such correspondence. In our example the two observations correspond to a particular email.

The variable that is suspected to be the response variable is plotted on the vertical (y) axis and the variable that is suspected to be the explanatory variable is plotted on the horizontal (x) axis. In this example, the variables could be switched since either variable could reasonably serve as the explanatory variable or the response variable.

DRAWING SCATTERPLOTS

(1) Decide which variable should go on each axis, and draw and label the two axes.
(2) Note the range of each variable, and add tick marks and scales to each axis.
(3) Plot the dots as you would on an (x, y) coordinate plane.

The association between two variables can be **positive** or **negative**, or there can be no association. Positive association means that larger values of the first variable are associated with larger values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.

EXAMPLE 2.2

What would it mean for two variables to have a *negative* association? What about *no* association?

Negative association implies that larger values of the first variable are associated with smaller values of the second variable. No association implies that the values of the second variable tend to be independent of changes in the first variable.

EXAMPLE 2.3

Figure 2.2 shows a plot of median household income against the poverty rate for 3,142 counties. What can be said about the relationship between these variables?

The relationship is evidently **nonlinear**, as highlighted by the dashed line. This is different from previous scatterplots we've seen, which show relationships that do not show much, if any, curvature in the trend. There is also a negative association, as higher rates of poverty tend to be associated with lower median household income.
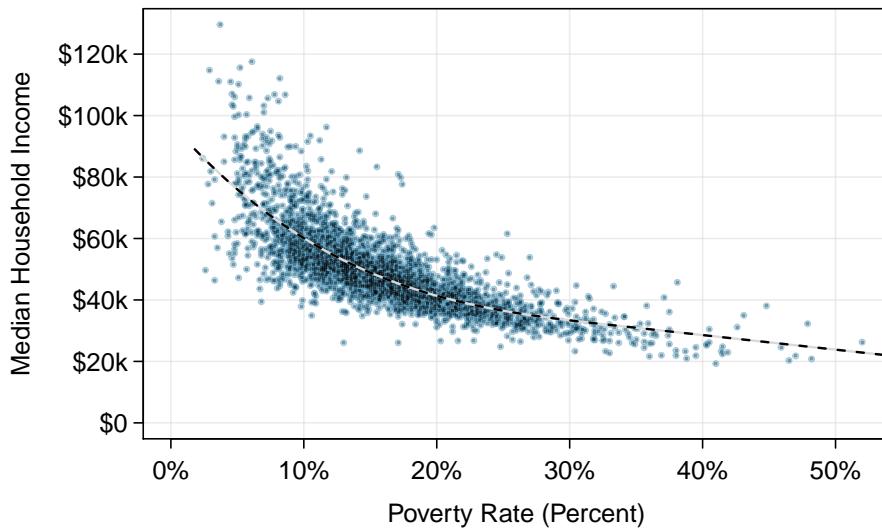
Figure 2.2: A scatterplot of the median household income against the poverty rate for the `county` data set. A statistical model has also been fit to the data and is shown as a dashed line. Explore dozens of scatterplots using American Community Survey data on Tableau Public ⁺👆.

GUIDED PRACTICE 2.4

What do scatterplots reveal about the data, and how are they useful?[1]

GUIDED PRACTICE 2.5

Describe two variables that would have a horseshoe-shaped association in a scatterplot (∩ or ∪).[2]

---

[1]Answers may vary. Scatterplots are helpful in quickly spotting associations relating variables, whether those associations come in the form of simple trends or whether those relationships are more complex.

[2]Consider the case where your vertical axis represents something "good" and your horizontal axis represents something that is only good in moderation. Health and water consumption fit this description: we require some water to survive, but consume too much and it becomes toxic and can kill a person. If health was represented on the vertical axis and water consumption on the horizontal axis, then we would create a ∩ shape.

2.1.2 Stem-and-leaf plots and dot plots

Sometimes two variables is one too many: only one variable may be of interest. In these cases we want to focus not on the association between two variables, but on the distribution of a single, or **univariate**, variable. The term **distribution** refers to the values that a variable takes and the frequency of these values. Here we introduce a new data set, the `email50` data set. This data set contains the number of characters in 50 emails. To simplify the data, we will round the numbers and record the values in thousands. Thus, 22105 is recorded as 22.

22	0	64	10	6	26	25	11	4	14
7	1	10	2	7	5	7	4	14	3
1	5	43	0	0	3	25	1	9	1
2	9	0	5	3	6	26	11	25	9
42	17	29	12	27	10	0	0	1	16

Figure 2.3: The number of characters, in thousands, for the data set of 50 emails.

Rather than look at the data as a list of numbers, which makes the distribution difficult to discern, we will organize it into a table called a **stem-and-leaf plot** shown in Figure 2.4. In a stem-and-leaf plot, each number is broken into two parts. The first part is called the **stem** and consists of the beginning digit(s). The second part is called the **leaf** and consists of the final digit(s). The stems are written in a column in ascending order, and the leaves that match up with those stems are written on the corresponding row. Figure 2.4 shows a stem-and-leaf plot of the number of characters in 50 emails. The stem represents the ten thousands place and the leaf represents the thousands place. For example, 1 | 2 corresponds to 12 thousand. When making a stem-and-leaf plot, remember to include a legend that describes what the stem and what the leaf represent. Without this, there is no way of knowing if 1 | 2 represents 1.2, 12, 120, 1200, etc.

```
0 | 00000011111223334455566777999
1 | 0001124467
2 | 25556679
3 |
4 | 23
5 |
6 | 4

Legend: 1 | 2 = 12,000
```

Figure 2.4: A stem-and-leaf plot of the number of characters in 50 emails.

G

GUIDED PRACTICE 2.6

There are a lot of numbers on the first row of the stem-and-leaf plot. Why is this the case?[3]
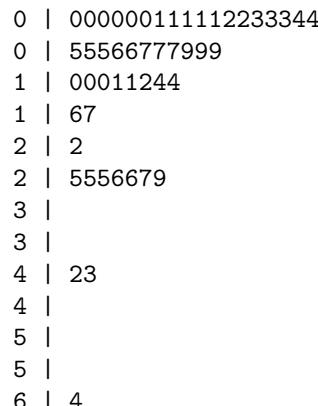
When there are too many numbers on one row or there are only a few stems, we *split* each row into two halves, with the leaves from 0-4 on the first half and the leaves from 5-9 on the second half. The resulting graph is called a **split stem-and-leaf plot**. Figure 2.5 shows the previous stem-and-leaf redone as a split stem-and-leaf.

G

GUIDED PRACTICE 2.7

What is the smallest number in the `email50` data set? What is the largest?[4]

---

[3]There are a lot of numbers on the first row because there are a lot of values in the data set less than 10 thousand.
[4]The smallest number is less than 1 thousand, and the largest is 64 thousand. That is a big range!

```
0 | 000000111112233344
0 | 55566777999
1 | 00011244
1 | 67
2 | 2
2 | 5556679
3 |
3 |
4 | 23
4 |
5 |
5 |
6 | 4

Legend: 1 | 2 = 12,000
```

Figure 2.5: A split stem-and-leaf.

Another simple graph for univariate numerical data is a dot plot. A **dot plot** uses dots to show the **frequency**, or number of occurrences, of the values in a data set. The higher the stack of dots, the greater the number occurrences there are of the corresponding value. An example using the same data set, number of characters from 50 emails, is shown in Figure 2.6.
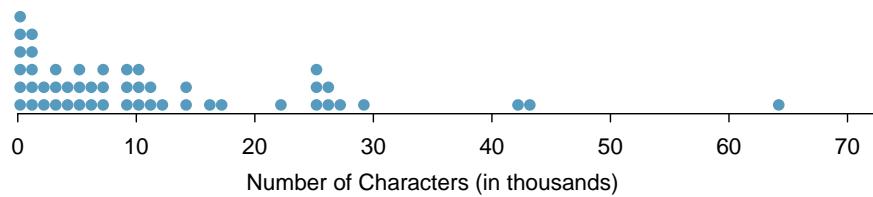
Figure 2.6: A dot plot of `num_char` for the `email50` data set.

GUIDED PRACTICE 2.8

Imagine rotating the dot plot 90 degrees clockwise. What do you notice?[5]

These graphs make it easy to observe important features of the data, such as the location of clusters and presence of gaps.

EXAMPLE 2.9

Based on both the stem-and-leaf and dot plot, where are the values clustered and where are the gaps for the `email50` data set?

———

There is a large cluster in the 0 to less than 20 thousand range, with a peak around 1 thousand. There are gaps between 30 and 40 thousand and between the two values in the 40 thousands and the largest value of approximately 64 thousand.

Additionally, we can easily identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called **outliers**. Later in this chapter we will provide numerical rules of thumb for identifying outliers. For now, it is sufficient to identify them by observing gaps in the graph. In this case, it would be reasonable to classify the emails with character counts of 42 thousand, 43 thousand, and 64 thousand as outliers since they are numerically distant from most of the data.

---

[5]It has a similar shape as the stem-and-leaf plot! The values on the horizontal axis correspond to the stems and the number of dots in each interval correspond the number of leaves needed for each stem.

OUTLIERS ARE EXTREME

An **outlier** is an observation that appears extreme relative to the rest of the data.

WHY IT IS IMPORTANT TO LOOK FOR OUTLIERS

Examination of data for possible outliers serves many useful purposes, including

1. Identifying asymmetry in the distribution.

2. Identifying data collection or entry errors. For instance, we re-examined the email purported to have 64 thousand characters to ensure this value was accurate.

3. Providing insight into interesting properties of the data.

GUIDED PRACTICE 2.10

The observation 64 thousand, a suspected outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[6]

GUIDED PRACTICE 2.11

Consider a data set that consists of the following numbers: 12, 12, 12, 12, 12, 13, 13, 14, 14, 15, 19. Which graph would better illustrate the data: a stem-and-leaf plot or a dot plot? Explain.[7]

2.1.3 Histograms

Stem-and-leaf plots and dot plots are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger samples. For larger samples, rather than showing the frequency of every value, we prefer to think of the value as belonging to a *bin*. For example, in the `email50` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Such a table, shown in Figure 2.7, is called a **frequency table**. Bins usually include the observations that fall on their left (lower) boundary and exclude observations that fall on their right (upper) boundary. This is called *left inclusive*. For example, 5 (i.e. 5000) would be counted in the 5-10 bin, not in the 0-5 bin. These binned counts are plotted as bars in Figure 2.8 into what is called a **histogram** or **frequency histogram**, which resembles the stacked dot plot shown in Figure 2.6.

Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	25-30	⋯	55-60	60-65
Count	19	12	6	2	3	5	⋯	0	1

Figure 2.7: The counts for the binned `num_char` data.

GUIDED PRACTICE 2.12

What can you see in the dot plot and stem-and-leaf plot that you cannot see in the frequency histogram?[8]

---

[6] That occasionally there may be very long emails.

[7] Because all the values begin with 1, there would be only one stem (or two in a split stem-and-leaf). This would not provide a good sense of the distribution. For example, the gap between 15 and 19 would not be visually apparent. A dot plot would be better here.

[8] Character counts for individual emails.
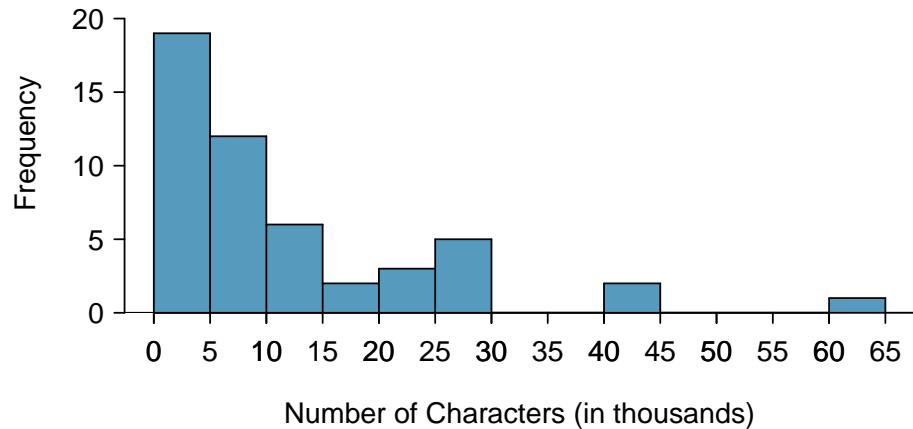
Figure 2.8: A histogram of `num_char`. This histogram uses bins or class intervals of width 5. Explore this histogram and dozens of histograms using American Community Survey data on Tableau Public.

DRAWING HISTOGRAMS

1. The variable is always placed on the horizontal axis. Before drawing the histogram, label both axes and draw a scale for each.

2. Draw bars such that the height of the bar is the frequency of that bin and the width of the bar corresponds to the bin width.

Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common. For instance, there are many more emails between 0 and 10,000 characters than emails between 10,000 and 20,000 in the data set. The bars make it easy to see how the density of the data changes relative to the number of characters.

EXAMPLE 2.13

How many emails had fewer than 10 thousand characters?

The height of the bars corresponds to frequency. There were 19 cases from 0 to less than 5 thousand and 12 cases from 5 thousand to less than 10 thousand, so there were $19 + 12 = 31$ emails with fewer than 10 thousand characters.

EXAMPLE 2.14

Approximately how many emails had fewer than 1 thousand chacters?

Based just on this histogram, we cannot know the exact answer to this question. We only know that 19 emails had between 0 and 5 thousand characters. If the number of emails is evenly distribution on this interval, then we can estimate that approximately $19/5 \approx 4$ emails fell in the range between 0 and 1 thousand.

EXAMPLE 2.15

What *percent* of the emails had 10 thousand or more characters?

From the first example, we know that 31 emails had fewer than 10 thousand characters. Since there are 50 emails in total, there must be 19 emails that have 10 thousand or more characters. To find the percent, compute $19/50 = 0.38 = 38\%$.

Sometimes questions such as the previous ones can be answered more easily with a **cumulative frequency histogram**. This type of histogram shows cumulative, or total, frequency achieved by each bin, rather than the frequency in that particular bin.

Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	25-30	30-35	$\cdots$	55-60	60-65
Cumulative Frequency	19	31	37	39	42	47	47	$\cdots$	49	50

Figure 2.9: The cumulative frequencies for the binned `num_char` data.
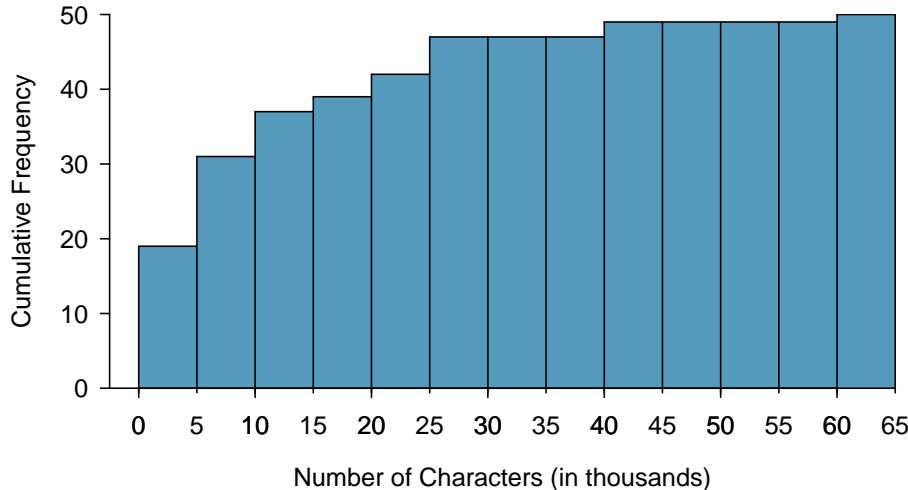
Figure 2.10: A cumulative frequency histogram of `num_char`. This histogram uses bins or class intervals of width 5. Compare frequency, relative frequency, cumulative frequency, and cumulative relative frequency histograms on Tableau Public ⁺↖.

EXAMPLE 2.16

How many of the emails had fewer than 20 thousand characters?

By tracing the height of the 15-20 thousand bin over to the vertical axis, we can see that it has a height just under 40 on the cumulative frequency scale. Therefore, we estimate that ≈39 of the emails had fewer than 30 thousand characters. Note that, unlike with a regular frequency histogram, we do not add up the height of the bars in a cumulative frequency histogram because each bar already represents a cumulative sum.

EXAMPLE 2.17

Using the cumulative frequency histogram, how many of the emails had 10-15 thousand characters?

To answer this question, we do a subtraction. ≈39 had fewer than 15-20 thousand emails and ≈37 had fewer than 10-15 thousand emails, so ≈2 must have had between 10-15 thousand emails.

EXAMPLE 2.18

Approximately 25 of the emails had fewer than how many characters?

This time we are given a cumulative frequency, so we start at 25 on the vertical axis and trace it across to see which bin it hits. It hits the 5-10 thousand bin, so 25 of the emails had fewer than a value somewhere between 5 and 10 thousand characters.

Knowing that 25 of the emails had fewer than a value between 5 and 10 thousand characters is useful information, but it is even more useful if we know what percent of the total 25 represents. Knowing that there were 50 total emails tells us that $25/50 = 0.5 = 50\%$ of the emails had fewer than a value between 5 and 10 thousand characters. When we want to know what fraction or percent of the data meet a certain criteria, we use relative frequency instead of frequency. **Relative frequency** is a fancy term for percent or proportion. It tells us how large a number is relative to the total.

Just as we constructed a frequency table, frequency histogram, and cumulative frequency histogram, we can construct a relative frequency table, relative frequency histogram, and cumulative relative frequency histogram.

GUIDED PRACTICE 2.19

How will the *shape* of the relative frequency histograms differ from the frequency histograms?[9]

PAY CLOSE ATTENTION TO THE VERTICAL AXIS OF A HISTOGRAM

We can misinterpret a histogram if we forget to check whether the vertical axis represents frequency, relative frequency, cumulative frequency, or cumulative relative frequency.

2.1.4   Describing Shape

Frequency and relative frequency histograms are especially convenient for describing the **shape** of the data distribution. Figure 2.8 shows that most emails have a relatively small number of characters, while fewer emails have a very large number of characters. When data trail off to the right in this way and have a longer right tail, the shape is said to be **right skewed**.[10]

Data sets with the reverse characteristic – a long, thin tail to the left – are said to be **left skewed**. We also say that such a distribution has a long left tail. Data sets that show roughly equal trailing off in both directions are called **symmetric**.

LONG TAILS TO IDENTIFY SKEW

When data trail off in one direction, the distribution has a **long tail**. If a distribution has a long left tail, it is left skewed. If a distribution has a long right tail, it is right skewed.

GUIDED PRACTICE 2.20

Take a look at the dot plot in Figure 2.6. Can you see the skew in the data? Is it easier to see the skew in the frequency histogram, the dot plot, or the stem-and-leaf plot?[11]

GUIDED PRACTICE 2.21

Would you expect the distribution of number of pets per household to be right skewed, left skewed, or approximately symmetric? Explain.[12]

---

[9]The shape will remain exactly the same. Changing from frequency to relative frequency involves dividing all the frequencies by the same number, so only the vertical scale (the numbers on the y-axis) change.

[10]Other ways to describe data that are right skewed: **skewed to the right**, **skewed to the high end**, or **skewed to the positive end**.

[11]The skew is visible in all three plots. However, it is not easily visible in the cumulative frequency histogram.

[12]We suspect most households would have 0, 1, or 2 pets but that a smaller number of households will have 3, 4, 5, or more pets, so there will be greater density over the small numbers, suggesting the distribution will have a long right tail and be right skewed.

In addition to looking at whether a distribution is skewed or symmetric, histograms, stem-and-leaf plots, and dot plots can be used to identify modes. A **mode** is represented by a prominent peak in the distribution.[13] There is only one prominent peak in the histogram of `num_char`.

Figure 2.11 shows histograms that have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that in Figure 2.8 there was one prominent peak in the unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.
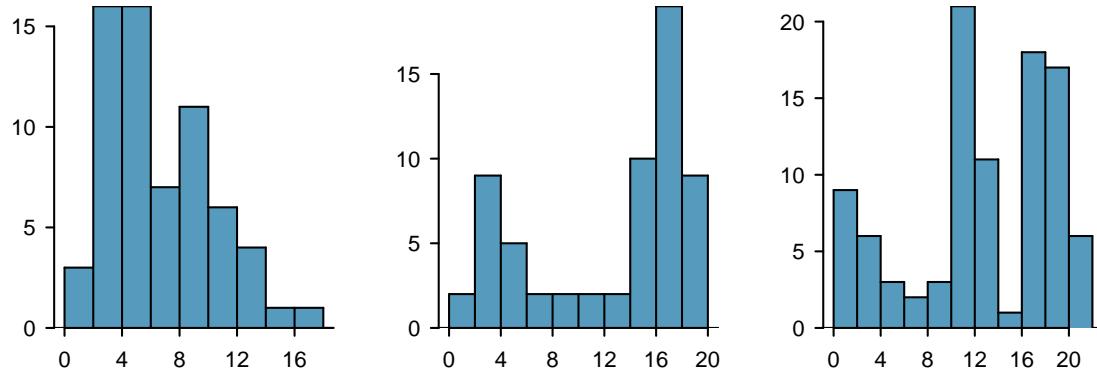
Figure 2.11: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

GUIDED PRACTICE 2.22

Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?[14]

LOOKING FOR MODES

Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why *prominent* is not rigorously defined in this book. The important part of this examination is to better understand your data and how it might be structured.

2.1.5 Descriptive versus inferential statistics

Finally, we note that the graphical summaries of this section and the numerical summaries of the next section fall into the realm of **descriptive statistics**. Descriptive statistics is about describing or summarizing data; it does not attribute properties of the data to a larger population. **Inferential statistics**, on the other hand, uses samples to generalize or to infer something about a larger population. We will have to wait until Chapter 5 to enter the exciting world of inferential statistics.

---

[13]Another definition of mode, which is not typically used in statistics, is the value with the most occurrences. It is common to have *no* observations with the same value in a data set, which makes this other definition useless for many real data sets.

[14]There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal.

Section summary

- A **scatterplot** is a **bivariate** display illustrating the relationship between two numerical variables. The observations must be **paired**, which is to say that they correspond to the same case or individual. The linear association between two variables can be positive or negative, or there can be no association. **Positive association** means that larger values of the first variable are associated with larger values of the second variable. **Negative association** means that larger values of the first variable are associated with smaller values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.

- When looking at a **univariate** display, researchers want to understand the distribution of the variable. The term **distribution** refers to the values that a variable takes and the frequency of those values. When looking at a distribution, note the presence of clusters, gaps, and **outliers**.

- Distributions may be **symmetric** or they may have a long tail. If a distribution has a long left tail (with greater density over the higher numbers), it is **left skewed**. If a distribution has a long right tail (with greater density over the smaller numbers), it is **right skewed**.

- Distributions may be **unimodal**, **bimodal**, or **multimodal**.

- Two graphs that are useful for showing the distribution of a small number of observations are the **stem-and-leaf plot** and **dot plot**. These graphs are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger data sets.

- For larger data sets it is common to use a **frequency histogram** or a **relative frequency histogram** to display the distribution of a variable. This requires choosing bins of an appropriate width.

- To see cumulative amounts, use a **cumulative frequency histogram**. A **cumulative relative frequency histogram** is ideal for showing **percentiles**.

- **Descriptive statistics** describes or summarizes data, while **inferential statistics** uses samples to generalize or infer something about a larger population.

Exercises

2.1 ACS, Part I. Each year, the US Census Bureau surveys about 3.5 million households with The American Community Survey (ACS). Data collected from the ACS have been crucial in government and policy decisions, helping to determine the allocation of federal and state funds each year. Some of the questions asked on the survey are about their income, age (in years), and gender. The table below contains this information for a random sample of 20 respondents to the 2012 ACS.[15]

	Income	Age	Gender		Income	Age	Gender
1	53,000	28	male	11	670	34	female
2	1600	18	female	12	29,000	55	female
3	70,000	54	male	13	44,000	33	female
4	12,800	22	male	14	48,000	41	male
5	1,200	18	female	15	30,000	47	female
6	30,000	34	male	16	60,000	30	male
7	4,500	21	male	17	108,000	61	male
8	20,000	28	female	18	5,800	50	female
9	25,000	29	female	19	50,000	24	female
10	42,000	33	male	20	11,000	19	male

(a) Create a scatterplot of income vs. age, and describe the relationship between these two variables.

(b) Now create two scatterplots: one for income vs. age for males and another for females.

(c) How, if at all, do the relationships between income and age differ for males and females?

2.2 MLB stats. A baseball team's success in a season is usually measured by their number of wins. In order to win, the team has to have scored more points (runs) than their opponent in any given game. As such, number of runs is often a good proxy for the success of the team. The table below shows number of runs, home runs, and batting averages for a random sample of 10 teams in the 2014 Major League Baseball season.[16]

	Team	Runs	Home runs	Batting avg.
1	Baltimore	705	211	0.256
2	Boston	634	123	0.244
3	Cincinnati	595	131	0.238
4	Cleveland	669	142	0.253
5	Detroit	757	155	0.277
6	Houston	629	163	0.242
7	Minnesota	715	128	0.254
8	NY Yankees	633	147	0.245
9	Pittsburgh	682	156	0.259
10	San Francisco	665	132	0.255

(a) Draw a scatterplot of runs vs. home runs.

(b) Draw a scatterplot of runs vs. batting averages.

(c) Are home runs or batting averages more strongly associated with number of runs? Explain your reasoning.

---

[15]United States Census Bureau. Summary File. 2012 American Community Survey. U.S. Census Bureau's American Community Survey Office, 2013. Web.

[16]ESPN: MLB Team Stats - 2014.

2.3   Fiber in your cereal. The Cereal FACTS report provides information on nutrition content of cereals as well as who they are targeted for (adults, children, families). We have selected a random sample of 20 cereals from the data provided in this report. Shown below are the fiber contents (percentage of fiber per gram of cereal) for these cereals.[17]

	Brand	Fiber %		Brand	Fiber %
1	Pebbles Fruity	0.0%	11	Cinnamon Toast Crunch	3.3%
2	Rice Krispies Treats	0.0%	12	Reese's Puffs	3.4%
3	Pebbles Cocoa	0.0%	13	Cheerios Honey Nut	7.1%
4	Pebbles Marshmallow	0.0%	14	Lucky Charms	7.4%
5	Frosted Rice Krispies	0.0%	15	Pebbles Boulders Chocolate PB	7.4%
6	Rice Krispies	3.0%	16	Corn Pops	9.4%
7	Trix	3.1%	17	Frosted Flakes Reduced Sugar	10.0%
8	Honey Comb	3.1%	18	Clifford Crunch	10.0%
9	Rice Krispies Gluten Free	3.3%	19	Apple Jacks	10.7%
10	Frosted Flakes	3.3%	20	Dora the Explorer	11.1%

(a) Create a stem and leaf plot of the distribution of the fiber content of these cereals.

(b) Create a dot plot of the fiber content of these cereals.

(c) Create a histogram and a relative frequency histogram of the fiber content of these cereals.

(d) What percent of cereals contain more than 7% fiber?

2.4   Sugar in your cereal. The Cereal FACTS report from Exercise 2.3 also provides information on sugar content of cereals. We have selected a random sample of 20 cereals from the data provided in this report. Shown below are the sugar contents (percentage of sugar per gram of cereal) for these cereals.
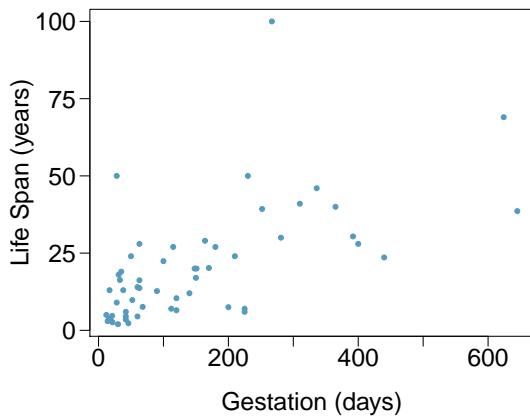
	Brand	Sugar %		Brand	Sugar %
1	Rice Krispies Gluten Free	3%	11	Corn Pops	31%
2	Rice Krispies	12%	12	Cheerios Honey Nut	32%
3	Dora the Explorer	22%	13	Reese's Puffs	34%
4	Frosted Flakes Red. Sugar	27%	14	Pebbles Fruity	37%
5	Clifford Crunch	27%	15	Pebbles Cocoa	37%
6	Rice Krispies Treats	30%	16	Lucky Charms	37%
7	Pebbles Boulders Choc. PB	30%	17	Frosted Flakes	37%
8	Cinnamon Toast Crunch	30%	18	Pebbles Marshmallow	37%
9	Trix	31%	19	Frosted Rice Krispies	40%
10	Honey Comb	31%	20	Apple Jacks	43%

(a) Create a stem and leaf plot of the distribution of the sugar content of these cereals.

(b) Create a dot plot of the sugar content of these cereals.

(c) Create a histogram and a relative frequency histogram of the sugar content of these cereals.

(d) What percent of cereals contain more than 30% sugar?

---
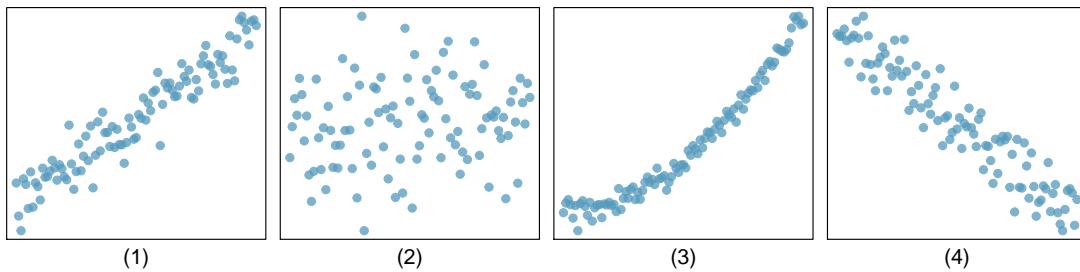
[17]JL Harris et al. "Cereal FACTS 2012: Limited progress in the nutrition quality and marketing of children's cereals". In: *Rudd Center for Food Policy & Obesity.* 12 (2012).

2.5 Mammal life spans. Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.[18]

(a) What type of an association is apparent between life span and length of gestation?

(b) What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

(c) Are life span and length of gestation independent? Explain your reasoning.

2.6 Associations. Indicate which of the plots show (a) a positive association, (b) a negative association, or (c) no association. Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.

(1)   (2)   (3)   (4)

[18]T. Allison and D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: *Arch. Hydrobiol* 75 (1975), p. 442.

2.2   Numerical summaries and box plots

What are the different ways to measure the center of a distribution, and why is there more than one way to measure the center? How do you know if a value is "far" from the center? What does it mean to an outlier? We will continue with the `email50` data set and investigate multiple quantitative summarizes for numerical data.

2.2.1   Learning objectives

1. Calculate, interpret, and compare the two measures of center (mean and median) and the three measures of spread (standard deviation, interquartile range, and range).
2. Understand how the shape of a distribution affects the relationship between the mean and the median.
3. Identify and apply the two rules of thumb for identify outliers (one involving standard deviation and mean and the other involving Q_1 and Q_3).
4. Describe the distribution a numerical variable with respect to center, spread, and shape, noting the presence of outliers.
5. Find the 5 number summary and IQR, and draw a box plot with outliers shown.
6. Understand the effect changing units has on each of the summary quantities.
7. Use the empirical rule to summarize approximately normal distributions.
8. Use quartiles, percentiles, and Z-scores to measure the relative position of a data point within the data set.
9. Compare the distribution of a numerical variable using dot plots / histograms with the same scale, back-to-back stem-and-leaf plots, or parallel box plots. Compare the distributions with respect to center, spread, shape, and outliers.

2.2.2   Measures of center

In the previous section, we saw that modes can occur anywhere in a data set. Therefore, mode is not a measure of **center**. We understand the term *center* intuitively, but quantifying what is the center can be a little more challenging. This is because there are different definitions of center. Here we will focus on the two most common: the mean and median.

The **mean**, sometimes called the average, is a common way to measure the center of a distribution of data. To find the mean number of characters in the 50 emails, we add up all the character counts and divide by the number of emails. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6$$

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `num_char`, and the bar on the x communicates that the average number of characters in the 50 emails was 11,600.

MEAN

The sample mean of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

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

where $\sum$ is the capital Greek letter sigma and $\sum x_i$ means take the sum of all the individual x values. $x_1, x_2, \ldots, x_n$ represent the n observed values.

GUIDED PRACTICE 2.23

Examine Equations (2.23) and (2.23) above. What does x_1 correspond to? And x_2? What does x_i represent?[19]

GUIDED PRACTICE 2.24

What was n in this sample of emails?[20]

The email50 data set represents a sample from a larger population of emails that were received in January and March. We could compute a mean for this population in the same way as the sample mean, however, the population mean has a special label: μ. The symbol μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x.

EXAMPLE 2.25

The average number of characters across all emails can be estimated using the sample data. Based on the sample of 50 emails, what would be a reasonable estimate of μ_x, the mean number of characters in all emails in the email data set? (Recall that email50 is a sample from email.)

The sample mean, 11,600, may provide a reasonable estimate of μ_x. While this number will not be perfect, it provides a *point estimate* of the population mean. In Chapter 5 and beyond, we will develop tools to characterize the reliability of point estimates, and we will find that point estimates based on larger samples tend to be more reliable than those based on smaller samples.

EXAMPLE 2.26

We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,142 counties in the county data set. What would be a better approach?

The county data set is special in that each county actually represents many individual people. If we were to simply average across the income variable, we would be treating counties with 5,000 and 5,000,000 residents equally in the calculations. Instead, we should compute the total income for each county, add up all the counties' totals, and then divide by the number of people in all the counties. If we completed these steps with the county data, we would find that the per capita income for the US is $27,348.43. Had we computed the *simple* mean of per capita income across counties, the result would have been just $22,504.70!

Example 2.26 used what is called a **weighted mean**, which will not be a key topic in this textbook. However, we have provided an online supplement on weighted means for interested readers:

www.openintro.org/stat/down/supp/wtdmean.pdf

---

[19] x_1 corresponds to the number of characters in the first email in the sample (21.7, in thousands), x_2 to the number of characters in the second email (7.0, in thousands), and x_i corresponds to the number of characters in the ith email in the data set.

[20] The sample size was $n = 50$.

The median provides another measure of center. The **median** splits an ordered data set in half. There are 50 character counts in the `email50` data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two middle observations: $(6{,}768 + 7{,}012)/2 = 6{,}890$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in this case that observation is the median (no average needed).

MEDIAN: THE NUMBER IN THE MIDDLE

In an ordered data set, the **median** is the observation right in the middle. If there are an even number of observations, the median is the average of the two middle values.

Graphically, we can think of the mean as the balancing point. The median is the value such that 50% of the *area* is to the left of it and 50% of the *area* is to the right of it.
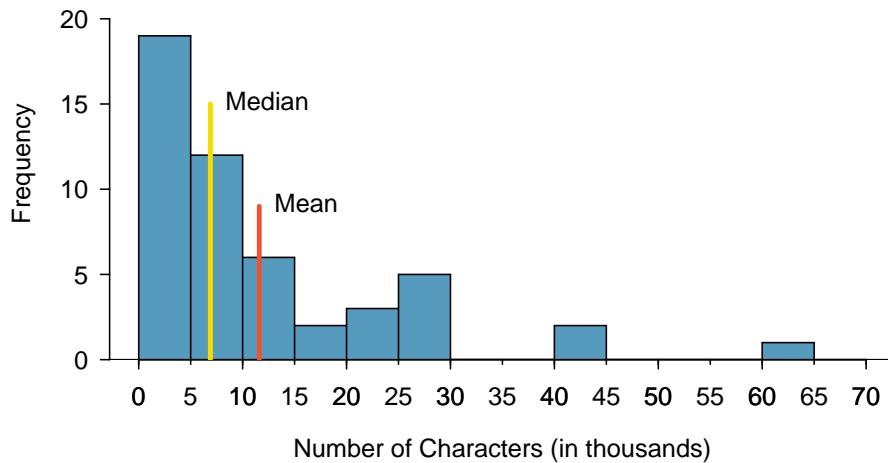
Figure 2.12: A histogram of `num_char` with its mean and median shown.

EXAMPLE 2.27

Based on the data, why is the mean greater than the median in this data set?

———————

Consider the three largest values of 42 thousand, 43 thousand, and 64 thousand. These values drag up the mean because they substantially increase the sum (the total). However, they do not drag up the median because their magnitude does not change the location of the middle value.

THE MEAN FOLLOWS THE TAIL

In a right skewed distribution, the mean is greater than the median.
In a left skewed distribution, the mean is less than the median.
In a symmetric distribution, the mean and median are approximately equal.

GUIDED PRACTICE 2.28

Consider the distribution of individual income in the United States. Which is greater: the mean or median? Why?[21]

---

[21]Because a small percent of individuals earn extremely large amounts of money while the majority earn a modest amount, the distribution is skewed to the right. Therefore, the mean is greater than the median.

2.2.3   Standard deviation as a measure of spread

The U.S. Census Bureau reported that in 2017, the median family income was \$73,891 and the mean family income was \$99,114.[22] Is a family income of \$60,000 far from the mean or somewhat close to the mean? In order to answer this question, it is not enough to know the center of the data set and its **range** (maximum value - minimum value). We must know about the variability of the data set within that range. Low variability or small spread means that the values tend to be more clustered together. High variability or large spread means that the values tend to be far apart.

EXAMPLE 2.29

Is it possible for two data sets to have the same range but different spread? If so, give an example. If not, explain why not.

Yes. An example is: 1, 1, 1, 1, 1, 9, 9, 9, 9, 9 and 1, 5, 5, 5, 5, 5, 5, 5, 5, 9.

The first data set has a larger spread because values tend to be farther away from each other while in the second data set values are clustered together at the mean.

Here, we introduce the standard deviation as a measure of spread. Though its formula is a bit tedious to calculate by hand, the standard deviation is very useful in data analysis and roughly describes how far away, on average, the observations are from the mean.

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1st, 2nd, 3rd, and 50th observations in the `num_char` variable. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$x_1 - \bar{x} = 21.7 - 11.6 = 10.1$$
$$x_2 - \bar{x} = 7.0 - 11.6 = -4.6$$
$$x_3 - \bar{x} = 0.6 - 11.6 = -11.0$$
$$\vdots$$
$$x_{50} - \bar{x} = 15.8 - 11.6 = 4.2$$

If we square these deviations and then take an average, the result is about equal to the sample **variance**, denoted by s^2:

$$s^2 = \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1}$$
$$= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49}$$
$$= 172.44$$

We divide by $n-1$, rather than dividing by n, when computing the variance; you need not worry about this mathematical nuance for the material in this textbook. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

The **standard deviation** is defined as the square root of the variance:

$$s = \sqrt{172.44} = 13.13$$

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of x may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n. The x subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

---

[22] https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_S1901

CALCULATING THE STANDARD DEVIATION

The standard deviation is the square root of the variance. It is roughly the "typical" distance of the observations from the mean.

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

The variance is useful for mathematical reasons, but the standard deviation is easier to interpret because it has the same units as the data set. The units for variance will be the units squared (e.g. meters^2). Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[23] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

THINKING ABOUT THE STANDARD DEVIATION

It is useful to think of the standard deviation as the "typical" or "average" distance that observations fall from the mean.

In Chapter 4, we encounter a bell-shaped distribution known as the *normal distribution*. The **empirical rule** tells us that for normal distributions, about 68% of the data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean. However, as seen in Figures 2.13 and 2.14, these percentages generally do not hold if the distribution is not bell-shaped.
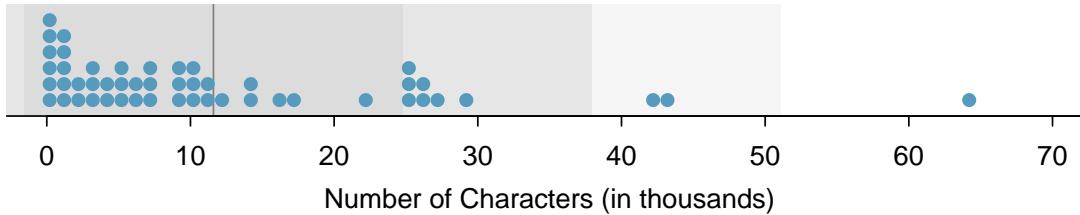
Figure 2.13: In the `num_char` data, 40 of the 50 emails (80%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. The empirical rule does not hold well for skewed data, as shown in this example.

GUIDED PRACTICE 2.30

On page 70, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.14 as an example, explain why such a description is important.[24]

---

[23]The only difference is that the population variance has a division by n instead of $n-1$.

[24]Figure 2.14 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.
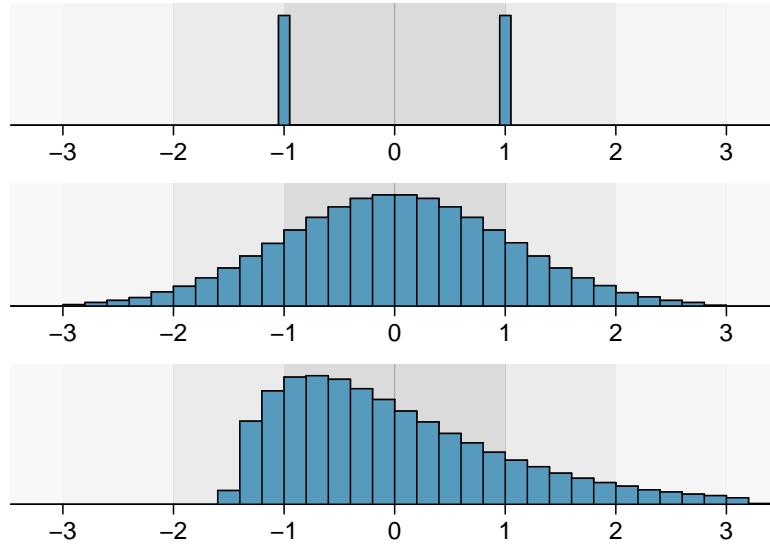
Figure 2.14: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

EXAMPLE 2.31

Earlier we reported that the mean family income in the U.S. in 2017 was $99,114. Estimating the standard deviation of income as approximately $50,000, is a family income of $60,000 far from the mean or relatively close to the mean?
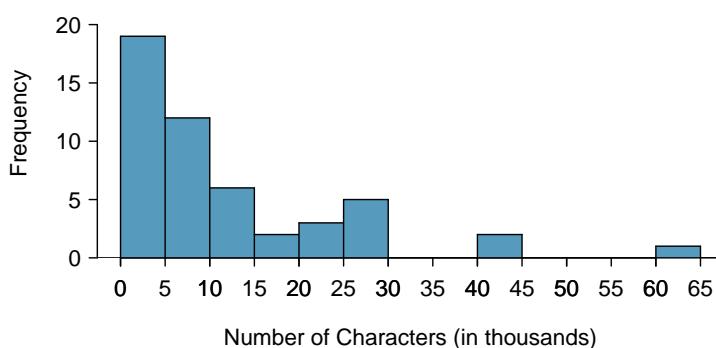
Because $60,000 is less that one standard deviation from the mean, it is relatively close to the mean. If the value were more than 2 standard deviations away from the mean, we would consider it far from the mean.

When describing any distribution, comment on the three important characteristics of center, spread, and shape. Also note any especially unusual cases.

EXAMPLE 2.32

In the data's context (the number of characters in emails), describe the distribution of the `num_char` variable shown in the histogram below.

The distribution of email character counts is unimodal and very strongly skewed to the right. Many of the counts fall near the mean at 11,600, and most fall within one standard deviation (13,130) of the mean. There is one exceptionally long email with about 65,000 characters.

In this chapter we use standard deviation as a descriptive statistic to describe the variability in a given data set. In Chapter 5 we will use the standard deviation to assess how close a sample mean is to the population mean.

2.2.4 Z-scores

Knowing how many standard deviations a value is from the mean is often more useful than simply knowing how far a value is from the mean.

EXAMPLE 2.33

Consider that the mean family income in the U.S. in 2017 was $99,114. Let's round this to $100,000 and estimate the standard deviation of income as $50,000. Using these estimates, how many standard deviations above the mean is an income of $200,000?

The value $200,000 is $100,000 above the mean. $100,000 is 2 standard deviations above the mean. This can be found by doing

$$\frac{200,000 - 100,000}{50,000} = 2$$

The number of standard deviations a value is above or below the mean is known as the **Z-score**. A Z-score has no units, and therefore is sometimes also called *standard units*.

THE Z-SCORE

The Z-score of an observation is the number of standard deviations it falls above or below the mean. We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

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

Observations above the mean always have positive Z-scores, while those below the mean always have negative Z-scores. If an observation is equal to the mean, then the Z-score is 0.

EXAMPLE 2.34

Head lengths of brushtail possums have a mean of 92.6 mm and standard deviation 3.6 mm. Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm.

For $x_1 = 95.4$ mm:

$$\begin{aligned} Z_1 &= \frac{x_1 - \mu}{\sigma} \\ &= \frac{95.4 - 92.6}{3.6} \\ &= 0.78 \end{aligned}$$

For $x_2 = 85.8$ mm:

$$\begin{aligned} Z_2 &= \frac{85.8 - 92.6}{3.6} \\ &= -1.89 \end{aligned}$$

We can use Z-scores to roughly identify which observations are more unusual than others. An observation x_1 is said to be more unusual than another observation x_2 if the absolute value of its Z-score is larger than the absolute value of the other observation's Z-score: $|Z_1| > |Z_2|$. This technique is especially insightful when a distribution is symmetric.

GUIDED PRACTICE 2.35

Which of the observations in Example 2.34 is more unusual?[25]

GUIDED PRACTICE 2.36

Let X represent a random variable from a distribution with $\mu = 3$ and $\sigma = 2$, and suppose we observe $x = 5.19$.
(a) Find the Z-score of x.
(b) Interpret the Z-score.[26]

Because Z-scores have no units, they are useful for comparing distance to the mean for distributions that have different standard deviations or different units.

EXAMPLE 2.37

The average daily high temperature in June in LA is 77°F with a standard deviation of 5°F. The average daily high temperature in June in Iceland is 13°C with a standard deviation of 3°C. Which would be considered more unusual: an 83°F day in June in LA or a 19°C day in June in Iceland?

———————

Both values are 6° above the mean. However, they are not the same number of standard deviations above the mean. 83 is $(83-77)/5 = 1.2$ standard deviations above the mean, while 19 is $(19-13)/3 = 2$ standard deviations above the mean. Therefore, a 19°C day in June in Iceland would be more unusual than an 83°F day in June in LA.

2.2.5 Box plots and quartiles

A **box plot** summarizes a data set using five summary statistics while also plotting unusual observations, called outliers. Figure 2.15 provides a box plot of the `num_char` variable from the `email50` data set.

The five summary statistics used in a box plot are known as the **five-number summary**, which consists of the minimum, the maximum, and the three quartiles (Q_1, Q_2, Q_3) of the data set being studied.

Q_2 represents the **second quartile**, which is equivalent to the 50th percentile (i.e. the median). Previously, we saw that Q_2 (the median) for the `email50` data set was the average of the two middle values: $\frac{6,768+7,012}{2} = 6,890$.

Q_1 represents the **first quartile**, which is the 25th percentile, and is the median of the smaller half of the data set. There are 25 values in the lower half of the data set, so Q_1 is the middle value: 2,454 characters. Q_3 represents the **third quartile**, or 75th percentile, and is the median of the larger half of the data set: 15,829 characters.

We calculate the variability in the data using the range of the middle 50% of the data: $Q_3 - Q_1 = 13,375$. This quantity is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability or **spread** in data. The more variable the data, the larger the standard deviation and IQR tend to be.

---

[25]Because the *absolute value* of Z-score for the second observation ($x_2 = 85.8$ mm $\rightarrow Z_2 = -1.89$) is larger than that of the first ($x_1 = 95.4$ mm $\rightarrow Z_1 = 0.78$), the second observation has a more unusual head length.

[26](a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$. (b) The observation x is 1.095 standard deviations *above* the mean. We know it must be above the mean since Z is positive.
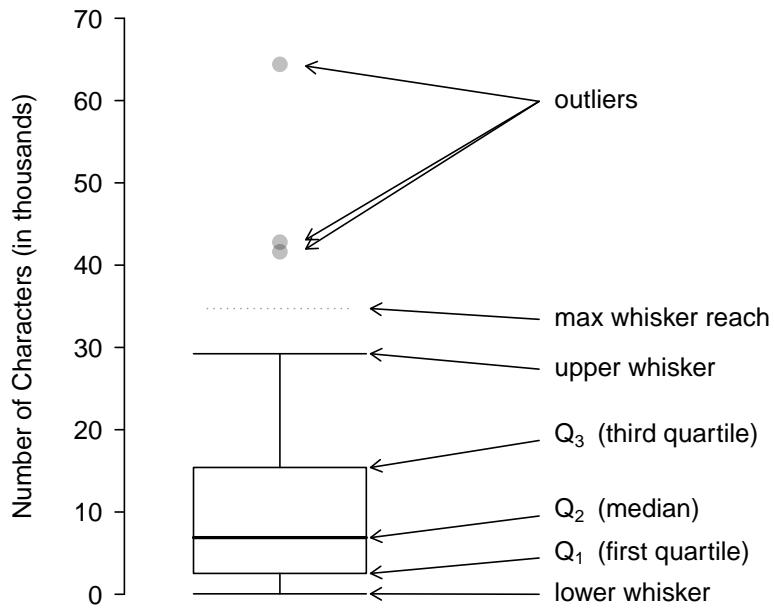
Figure 2.15: A labeled box plot for the number of characters in 50 emails. The median (6,890) splits the data into the bottom 50% and the top 50%. Explore dozens of boxplots with histograms using American Community Survey data on Tableau Public ⁺↖.

INTERQUARTILE RANGE (IQR)

The IQR is the length of the box in a box plot. It is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25th and 75th percentiles.

OUTLIERS IN THE CONTEXT OF A BOX PLOT

When in the context of a box plot, define an **outlier** as an observation that is more than $1.5 \times IQR$ above Q_3 or $1.5 \times IQR$ below Q_1. Such points are marked using a dot or asterisk in a box plot.

To build a box plot, draw an axis (vertical or horizontal) and draw a scale. Draw a dark line denoting Q_2, the median. Next, draw a line at Q_1 and at Q_3. Connect the Q_1 and Q_3 lines to form a rectangle. The width of the rectangle corresponds to the IQR and the middle 50% of the data is in this interval.

Extending out from the rectangle, the **whiskers** attempt to capture all of the data remaining outside of the box, except outliers. In Figure 2.15, the upper whisker does not extend to the last three points, which are beyond $Q_3 + 1.5 \times IQR$ and are outliers, so it extends only to the last point below this limit.[27] The lower whisker stops at the lowest value, 33, since there are no additional data to reach. Outliers are each marked with a dot or asterisk. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

---

[27]You might wonder, isn't the choice of $1.5 \times IQR$ for defining an outlier arbitrary? It is! In practical data analyses, we tend to avoid a strict definition since what is an unusual observation is highly dependent on the context of the data.

EXAMPLE 2.38

Compare the box plot to the graphs previously discussed: stem-and-leaf plot, dot plot, frequency and relative frequency histogram. What can we learn more easily from a box plot? What can we learn more easily from the other graphs?

———————

It is easier to immediately identify the quartiles from a box plot. The box plot also more prominently highlights outliers. However, a box plot, unlike the other graphs, does not show the *distribution* of the data. For example, we cannot generally identify modes using a box plot.

EXAMPLE 2.39

Is it possible to identify skew from the box plot?

———————

Yes. Looking at the lower and upper whiskers of this box plot, we see that the lower 25% of the data is squished into a shorter distance than the upper 25% of the data, implying that there is greater density in the low values and a tail trailing to the upper values. This box plot is right skewed.

GUIDED PRACTICE 2.40

True or false: there is more data between the median and Q_3 than between Q_1 and the median.[28]

EXAMPLE 2.41

Consider the following ordered data set.

$$5 \quad 5 \quad 9 \quad 10 \quad 15 \quad 16 \quad 20 \quad 30 \quad 80$$

Find the 5 number summary and identify how small or large a value would need to be to be considered an outlier. Are there any outliers in this data set?

———————

There are nine numbers in this data set. Because n is odd, the median is the middle number: 15. When finding Q_1, we find the median of the lower half of the data, which in this case includes 4 numbers (we do not include the 15 as belonging to either half of the data set). Q_1 then is the average of 5 and 9, which is $Q_1 = 7$, and Q_3 is the average of 20 and 30, so $Q_3 = 25$. The min is 5 and the max is 80. To see how small a number needs to be to be an outlier on the low end we do:

$$\begin{aligned} Q_1 - 1.5 \times IQR &= Q_1 - 1.5 \times (Q_3 - Q_1) \\ &= 7 - 1.5 \times (35 - 7) \\ &= -35 \end{aligned}$$

On the high end we need:

$$\begin{aligned} Q_3 + 1.5 \times IQR &= Q_3 + 1.5 \times (Q_3 - Q_1) \\ &= 35 + 1.5 \times (35 - 7) \\ &= 77 \end{aligned}$$

There are no numbers less than -41, so there are no outliers on the low end. The observation at 80 is greater than 77, so 80 is an outlier on the high end.

———————————————————

[28]False. Since Q_1 is the 25th percentile and the median is the 50th percentile, 25% of the data fall between Q_1 and the median. Similarly, 25% of the data fall between Q_2 and the median. The distance between the median and Q_3 is larger because that 25% of the data is more spread out.

2.2.6   Calculator/Desmos: summarizing 1-variable statistics

One can use a handheld calculator or online software such as Desmos to calculate summary statistics. More advanced statistical software packages include R (in which most of the graphs in this text were made), Python, SAS, and STATA.

Email50 Data Set

1-Var Stats

$\bar{x} = 11.6$
$s_x = 13.13$
$\sigma_x = 12.99$
$n = 50$
$\min = 0.057$
$Q_1 = 2.454$
$med = 6.8895$
$Q_3 = 15.829$
$\max = 64.401$
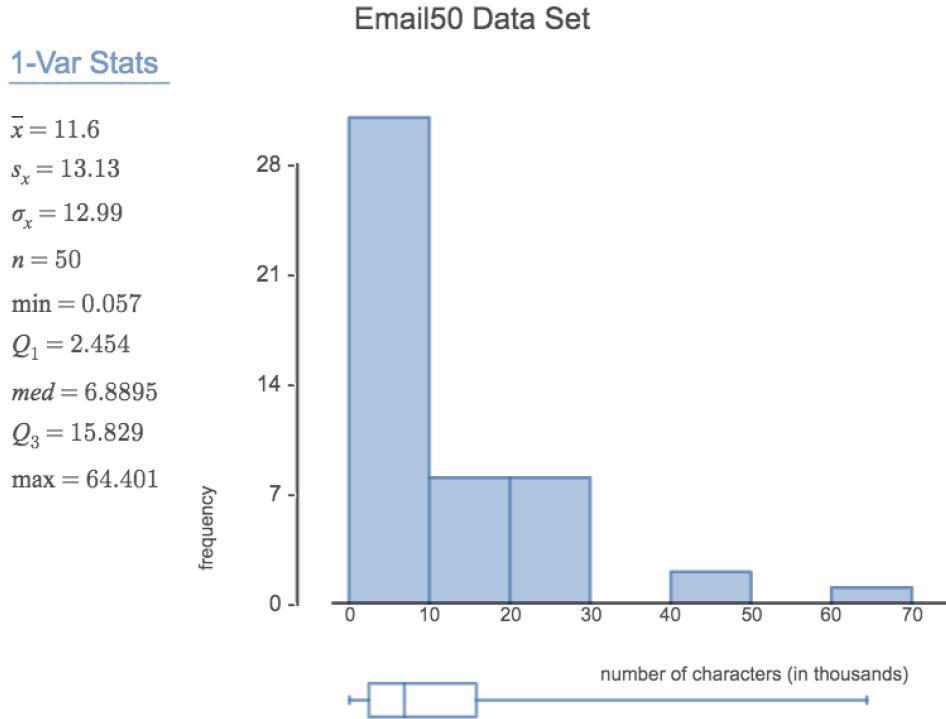
Figure 2.16: Use this 1-Var Stats calculator (openintro.org/ahss/desmos) to graph and find summary statistics for a single variable in Desmos, as shown in the figure.

TI-83/84: ENTERING DATA

The first step in summarizing data or making a graph is to enter the data set into a list. Use STAT, Edit.

1. Press STAT.
2. Choose 1:Edit.
3. Enter data into L1 or another list.

CASIO FX-9750GII: ENTERING DATA

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. Optional: use the left or right arrows to select a particular list.
3. Enter each numerical value and hit EXE.

TI-84: CALCULATING SUMMARY STATISTICS

Use the `STAT`, `CALC`, `1-Var Stats` command to find summary statistics such as mean, standard deviation, and quartiles.

1. Enter the data as described previously.

2. Press `STAT`.

3. Right arrow to `CALC`.

4. Choose `1:1-Var Stats`.

5. Enter `L1` (i.e. `2ND 1`) for List. If the data is in a list other than `L1`, type the name of that list.

6. Leave `FreqList` blank.

7. Choose `Calculate` and hit `ENTER`.

TI-83: Do steps 1-4, then type `L1` (i.e. `2nd 1`) or the list's name and hit `ENTER`.

Calculating the summary statistics will return the following information. It will be necessary to hit the down arrow to see all of the summary statistics.

$\bar{x}$	Mean	`n`	Sample size or # of data points
Σx	Sum of all the data values	`minX`	Minimum
Σx^2	Sum of all the squared data values	Q_1	First quartile
`Sx`	Sample standard deviation	`Med`	Median
σx	Population standard deviation	`maxX`	Maximum

TI-83/84: DRAWING A BOX PLOT

1. Enter the data to be graphed as described previously.

2. Hit `2ND Y=` (i.e. `STAT PLOT`).

3. Hit `ENTER` (to choose the first plot).

4. Hit `ENTER` to choose `ON`.

5. Down arrow and then right arrow three times to select box plot with outliers.

6. Down arrow again and make `Xlist: L1` and `Freq: 1`.

7. Choose `ZOOM` and then `9:ZoomStat` to get a good viewing window.

CASIO FX-9750GII: DRAWING A BOX PLOT AND 1-VARIABLE STATISTICS

1. Navigate to STAT (MENU, then hit 2) and enter the data into a list.

2. Go to GRPH (F1).

3. Next go to SET (F6) to set the graphing parameters.

4. To use the 2nd or 3rd graph instead of GPH1, select F2 or F3.

5. Move down to Graph Type and select the ▷ (F6) option to see more graphing options, then select Box (F2).

6. If XList does not show the list where you entered the data, hit LIST (F1) and enter the correct list number.

7. Leave Frequency at 1.

8. For Outliers, choose On (F1).

9. Hit EXE and then choose the graph where you set the parameters F1 (most common), F2, or F3.

10. If desired, explore 1-variable statistics by selecting 1-Var (F1).

EXAMPLE 2.42
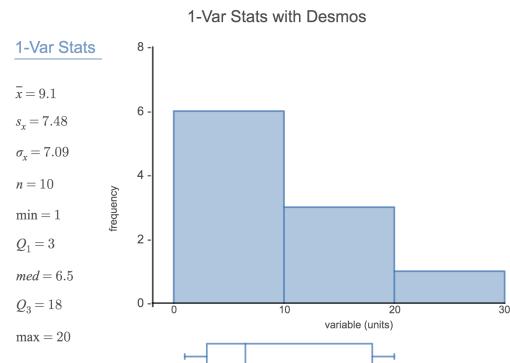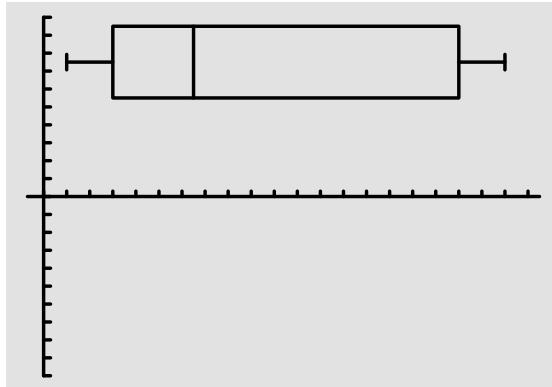
Enter the following 10 data points into a calculator or into this Desmos 1-Var Stats calculator (openintro.org/ahss/desmos):

$$5, 8, 1, 19, 3, 1, 11, 18, 20, 5$$

Find the summary statistics and make a box plot of the data.

The summary statistics should be $\bar{x} = 9.1$, $Sx = 7.48$, $Q1 = 3$, etc. The box plot should be as follows.

TI-83/84: WHAT TO DO IF YOU CANNOT FIND L1 OR ANOTHER LIST

Restore lists L1-L6 using the following steps:

1. Press STAT.

2. Choose 5:SetUpEditor.

3. Hit ENTER.

CASIO FX-9750GII: DELETING A DATA LIST

1. Navigate to STAT (MENU, then hit 2).
2. Use the arrow buttons to navigate to the list you would like to delete.
3. Select ▷ (F6) to see more options.
4. Select DEL-A (F4) and then F1 to confirm.

2.2.7 Outliers and robust statistics

RULES OF THUMB FOR IDENTIFYING OUTLIERS

There are two rules of thumb for identifying outliers:

- More than $1.5\times$ IQR below Q_1 or above Q_3
- More than 2 standard deviations above or below the mean.

Both are important for the AP exam. In practice, consider these to be only rough guidelines.

GUIDED PRACTICE 2.43

For the email50 data set, $Q_1 = 2{,}536$ and $Q_3 = 15{,}411$. $\bar{x} = 11{,}600$ and $s = 13{,}130$. What values would be considered an outlier on the low end using each rule?[29]

GUIDED PRACTICE 2.44

Because there are no negative values in this data set, there can be no outliers on the low end. What does the fact that there are outliers on the high end but not on the low end suggestion?[30]

How are the sample statistics of the num_char data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these summary statistics if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 2.17, and sample statistics are computed under each scenario in Figure 2.18.
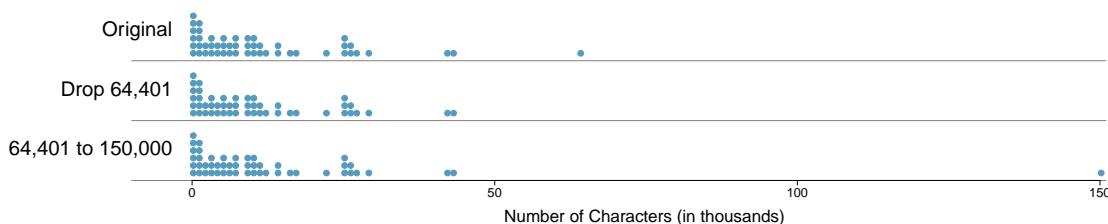
Figure 2.17: Dot plots of the original character count data and two modified data sets.

---

[29] $Q_1 - 1.5 \times IQR = 2536 - 1.5 \times (15411 - 2536) = -16{,}749.5$, so values less than -16,749.5 would be considered an outlier using the first rule of thumb. Using the second rule of thumb, a value less than $\bar{x} - 2 \times s = 11{,}600 - 2 \times 13{,}130 = -14{,}660$ would be considered an outlier. Note tht these are just rules of thumb and yield different values.

[30] It suggests that the distribution has a right hand tail, that is, that it is right skewed.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original num_char data	6,890	12,875	11,600	13,130
drop 64,401 observation	6,768	11,702	10,521	10,798
move 64,401 to 150,000	6,890	12,875	13,310	22,434

Figure 2.18: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

GUIDED PRACTICE 2.45

(a) Which is more affected by extreme observations, the mean or median? Figure 2.18 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[31]

The median and IQR are called **robust estimates** because extreme observations have little effect on their values. The mean and standard deviation are much more affected by changes in extreme observations.

EXAMPLE 2.46

The median and IQR do not change much under the three scenarios in Figure 2.18. Why might this be the case?

Since there are no large gaps between observations around the three quartiles, adding, deleting, or changing one value, no matter how extreme that value, will have little effect on their values.

GUIDED PRACTICE 2.47

The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[32]

2.2.8   Linear transformations of data

EXAMPLE 2.48

Begin with the following list: 1, 1, 5, 5. Multiply all of the numbers by 10. What happens to the mean? What happens to the standard deviation? How do these compare to the mean and the standard deviation of the original list?

The original list has a mean of 3 and a standard deviation of 2. The new list: 10, 10, 50, 50 has a mean of 30 with a standard deviation of 20. Because all of the values were multiplied by 10, both the mean and the standard deviation were multiplied by 10. [33]

---

[31](a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 2.45.

[32]Buyers of a "regular car" should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

[33]Here, the population standard deviation was used in the calculation. These properties can be proven mathematically using properties of sigma (summation).

EXAMPLE 2.49

Start with the following list: 1, 1, 5, 5. Multiply all of the numbers by -0.5. What happens to the mean? What happens to the standard deviation? How do these compare to the mean and the standard deviation of the original list?

—————

The new list: -0.5, -0.5, -2.5, -2.5 has a mean of -1.5 with a standard deviation of 1. Because all of the values were multiplied by -0.5, the mean was multiplied by -0.5. Multiplying all of the values by a negative flipped the sign of numbers, which affects the location of the center, but not the spread. Multiplying all of the values by -0.5 multiplied the standard deviation by +0.5 since the standard deviation cannot be negative.

EXAMPLE 2.50

Again, start with the following list: 1, 1, 5, 5. Add 100 to every entry. How do the new mean and standard deviation compare to the original mean and standard deviation?

—————

The new list is: 101, 101, 105, 105. The new mean of 103 is 100 greater than the original mean of 3. The new standard deviation of 2 is the *same* as the original standard deviation of 2. Adding a constant to every entry shifted the values, but did not stretch them.

Suppose that a researcher is looking at a list of 500 temperatures recorded in Celsius (C). The mean of the temperatures listed is given as 27°C with a standard deviation of 3°C. Because she is not familiar with the Celsius scale, she would like to convert these summary statistics into Fahrenheit (F). To convert from Celsius to Fahrenheit, we use the following conversion:

$$x_F = \frac{9}{5}x_C + 32$$

Fortunately, she does not need to convert each of the 500 temperatures to Fahrenheit and then re-calculate the mean and the standard deviation. The unit conversion above is a linear transformation of the following form, where $a = 9/5$ and $b = 32$:

$$aX + b$$

Using the examples as a guide, we can solve this temperature-conversion problem. The mean was 27°C and the standard deviation was 3°C. To convert to Fahrenheit, we multiply all of the values by 9/5, which multiplies both the mean and the standard deviation by 9/5. Then we add 32 to all of the values which adds 32 to the mean but does not change the standard deviation further.

$$\bar{x}_F = \frac{9}{5}\bar{x}_C + 32 \qquad\qquad \sigma_F = \frac{9}{5}\sigma_C$$
$$= \frac{5}{9}(27) + 32 \qquad\qquad = \frac{9}{5}(3)$$
$$= 80.6 \qquad\qquad\qquad = 5.4$$
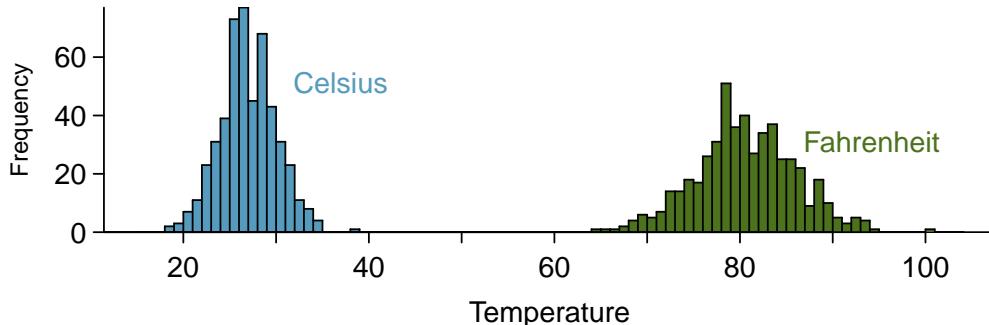
Figure 2.19: 500 temperatures shown in both Celsius and Fahrenheit.

ADDING SHIFTS THE VALUES, MULTIPLYING STRETCHES OR CONTRACTS THEM

Adding a constant to every value in a data set shifts the mean but does not affect the standard deviation. Multiplying the values in a data set by a constant will change the mean and the standard deviation by the same multiple, except that the standard deviation will always remain positive.

EXAMPLE 2.51

Consider the temperature example. How would converting from Celsuis to Fahrenheit affect the median? The IQR?

Ⓔ

The median is affected in the same way as the mean and the IQR is affected in the same way as the standard deviation. To get the new median, multiply the old median by 9/5 and add 32. The IQR is computed by subtracting Q_1 from Q_3. While Q_1 and Q_3 are each affected in the same way as the median, the additional 32 added to each will cancel when we take $Q_3 - Q_1$. That is, the IQR will be increase by a factor of 9/5 but will be unaffected by the addition of 32.

For a more mathematical explanation of the IQR calculation, see the footnote.[a]

---

[a]new IQR $= \left(\frac{9}{5}Q_3 + 32\right) - \left(\frac{9}{5}Q_1 + 32\right) = \frac{9}{5}\left(Q_3 - Q_1\right) = \frac{9}{5} \times$ (old IQR).

2.2.9 Comparing numerical data across groups

Some of the more interesting investigations can be considered by examining numerical data across groups. The methods required here aren't really new. All that is required is to make a numerical plot for each group. To make a direct comparison between two groups, create a pair of dot plots or a pair of histograms drawn using the same scales. It is also common to use back-to-back stem-and-leaf plots, parallel box plots, and hollow histograms, the three of which are explored here.

We will take a look again at the `county` data set and compare the median household income for counties that gained population from 2010 to 2017 versus counties that had no gain. While we might like to make a causal connection here, remember that these are observational data and so such an interpretation would be, at best, half-baked.

There were 1,454 counties where the population increased from 2010 to 2017, and there were 1,672 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Figure 2.20 to give a better sense of some of the raw median income data.

The **side-by-side box plot** is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 2.22, where there are two box plots, one for each group, placed into one plotting window and drawn on the same scale.

Another useful plotting method uses **hollow histograms** to compare numerical data across groups. These are just the outlines of histograms of each group put on the same plot, as shown in the right panel of Figure 2.22.

GUIDED PRACTICE 2.52

Ⓖ

Use the plots in Figure 2.22 to compare the incomes for counties across the two groups. What do you notice about the approximate center of each group? What do you notice about the variability between groups? Is the shape relatively consistent between groups? How many *prominent* modes are there for each group?[34]

---

[34]Answers may vary a little. The counties with population gains tend to have higher income (median of about $45,000) versus counties without a gain (median of about $40,000). The variability is also slightly larger for the population gain group. This is evident in the IQR, which is about 50% bigger in the *gain* group. Both distributions show slight to moderate right skew and are unimodal. The box plots indicate there are many observations far above the median in each group, though we should anticipate that many observations will fall beyond the whiskers when examining any data set that contain more than a couple hundred data points.

Median Income for 150 Counties, in $1000s

Population Gain						No Population Gain		
38.2	43.6	42.2	61.5	51.1	45.7	48.3	60.3	50.7
44.6	51.8	40.7	48.1	56.4	41.9	39.3	40.4	40.3
40.6	63.3	52.1	60.3	49.8	51.7	57	47.2	45.9
51.1	34.1	45.5	52.8	49.1	51	42.3	41.5	46.1
80.8	46.3	82.2	43.6	39.7	49.4	44.9	51.7	46.4
75.2	40.6	46.3	62.4	44.1	51.3	29.1	51.8	50.5
51.9	34.7	54	42.9	52.2	45.1	27	30.9	34.9
61	51.4	56.5	62	46	46.4	40.7	51.8	61.1
53.8	57.6	69.2	48.4	40.5	48.6	43.4	34.7	45.7
53.1	54.6	55	46.4	39.9	56.7	33.1	21	37
63	49.1	57.2	44.1	50	38.9	52	31.9	45.7
46.6	46.5	38.9	50.9	56	34.6	56.3	38.7	45.7
74.2	63	49.6	53.7	77.5	60	56.2	43	21.7
63.2	47.6	55.9	39.1	57.8	42.6	44.5	34.5	48.9
50.4	49	45.6	39	38.8	37.1	50.9	42.1	43.2
57.2	44.7	71.7	35.3	100.2		35.4	41.3	33.6
42.6	55.5	38.6	52.7	63		43.4	56.5	

Figure 2.20: In this table, median household income (in $1000s) from a random sample of 100 counties that had population gains are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

```
         Population: Gain          Population: No Gain


                             |  2 |12
                             |  2 |79
                           4|  3 |1234
                  99999987555|  3 |5555799
               444433322111000|  4 |00111223333
        99999888766666665555|  4 |55666666789
     444333222221111110000|  5 |1112222
              887776666555|  5 |6677
                33333222100|  6 |01
                         9|  6 |
                        42|  7 |
                        85|  7 |
                        21|  8 |


        Legend: 2 |1 = 21,000 median income
```
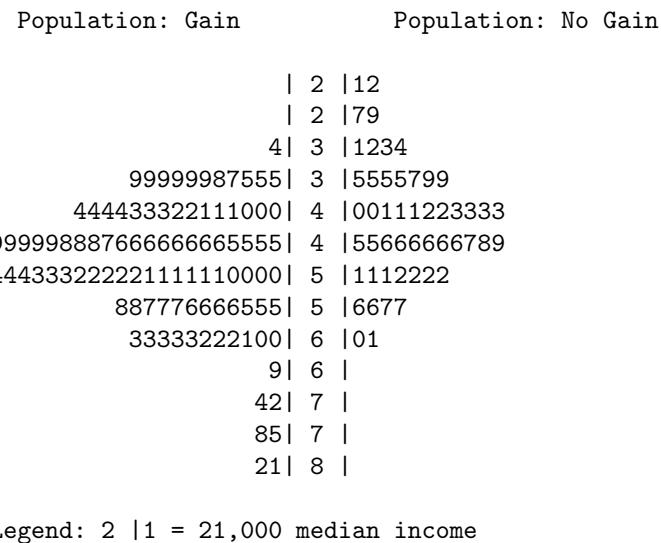
Figure 2.21: Back-to-back stem-and-leaf plot for median income, split by whether the count had a population gain or no gain.
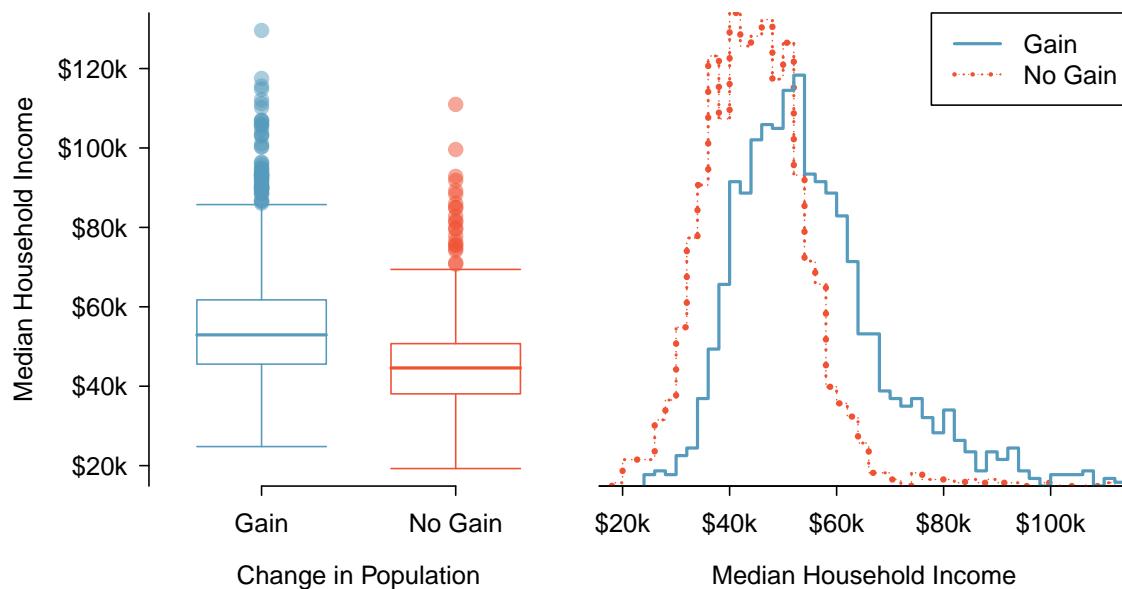
Figure 2.22: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_hh_income`, where the counties are split by whether or not there was a population gain from 2010 to 2017. Explore this data set on Tableau Public ⤢.

COMPARING DISTRIBUTIONS

When comparing distributions, compare them with respect to center, spread, and shape as well as any unusual observations. Such descriptions should be in context.

GUIDED PRACTICE 2.53

What components of each plot in Figure 2.22 do you find most useful?[35]

GUIDED PRACTICE 2.54

Do these graphs tell us about any association between income for the two groups?[36]

Looking at an association is different than comparing distributions. When comparing distributions, we are interested in questions such as, "Which distribution has a greater average?" and "How do the shapes of the distribution differ?" The number of elements in each data set need not be the same (e.g. height of women and height of men). When we look at association, we are interested in whether there is a positive, negative, or no association between the variables. This requires two data sets of equal length that are essentially paired (e.g. height and weight of individuals).

COMPARING DISTRIBUTIONS VERSUS LOOKING AT ASSOCIATION

We compare two distributions with respect to center, spread, and shape. To compare the distributions visually, we use 2 single-variable graphs, such as two histograms, two dot plots, parallel box plots, or a back-to-back stem-and-leaf. When looking at association, we look for a positive, negative, or no relationship between the variables. To see association visually, we require a scatterplot.

---

[35]Answers will vary. The parallel box plots are especially useful for comparing centers and spreads, while the hollow histograms are more useful for seeing distribution shape, skew, and groups of anomalies.

[36]No, to see association we require a scatterplot. Moreover, these data are not paired, so the discussion of association does not make sense here.

2.2.10 Mapping data (special topic)

The `county` data set offers many numerical variables that we could plot using dot plots, scatterplots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should create an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 2.23 and 2.24 shows intensity maps for poverty rate in percent (`poverty`), unemployment rate (`unemployment_rate`), homeownership rate in percent (`homeownership`), and median household income (`median_hh_income`). The color key indicates which colors correspond to which values. The intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions or hypotheses.

EXAMPLE 2.55

What interesting features are evident in the `poverty` and `unemployment_rate` intensity maps?

Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does much of Arizona and New Mexico. High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky.

The unemployment rate follows similar trends, and we can see correspondence between the two variables. In fact, it makes sense for higher rates of unemployment to be closely related to poverty rates. One observation that stand out when comparing the two maps: the poverty rate is much higher than the unemployment rate, meaning while many people may be working, they are not making enough to break out of poverty.

GUIDED PRACTICE 2.56

What interesting features are evident in the `median_hh_income` intensity map in Figure 2.24(b)?[37]

---
[37]Note: answers will vary. There is some correspondence between high earning and metropolitan areas, where we can see darker spots (higher median household income), though there are several exceptions. You might look for large cities you are familiar with and try to spot them on the map as dark spots.

(a)

(b)
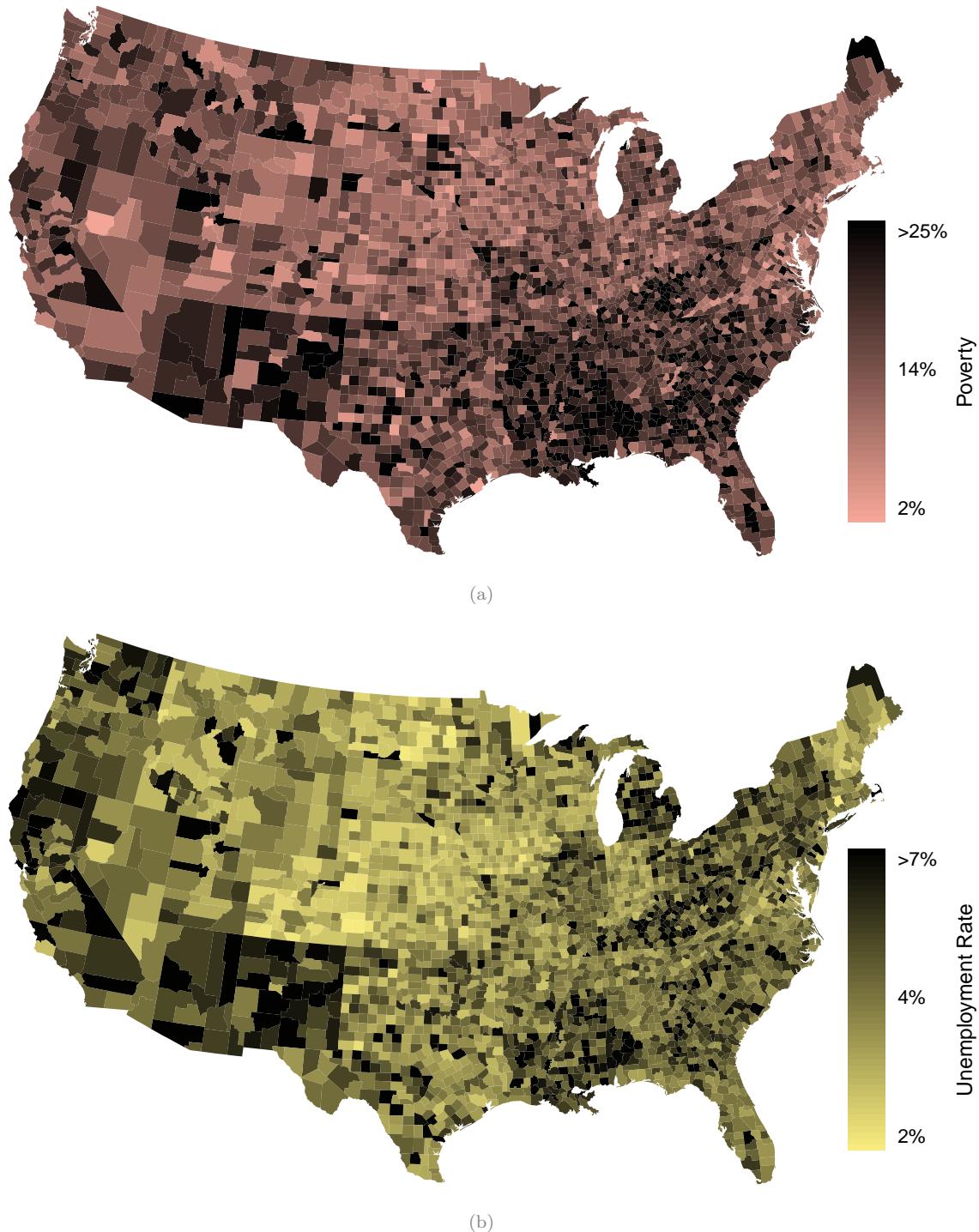
Figure 2.23: (a) Intensity map of poverty rate (percent). (b) Intensity map of the unemployment rate (percent). Explore dozens of intensity maps using American Community Survey data on Tableau Public ⁺.
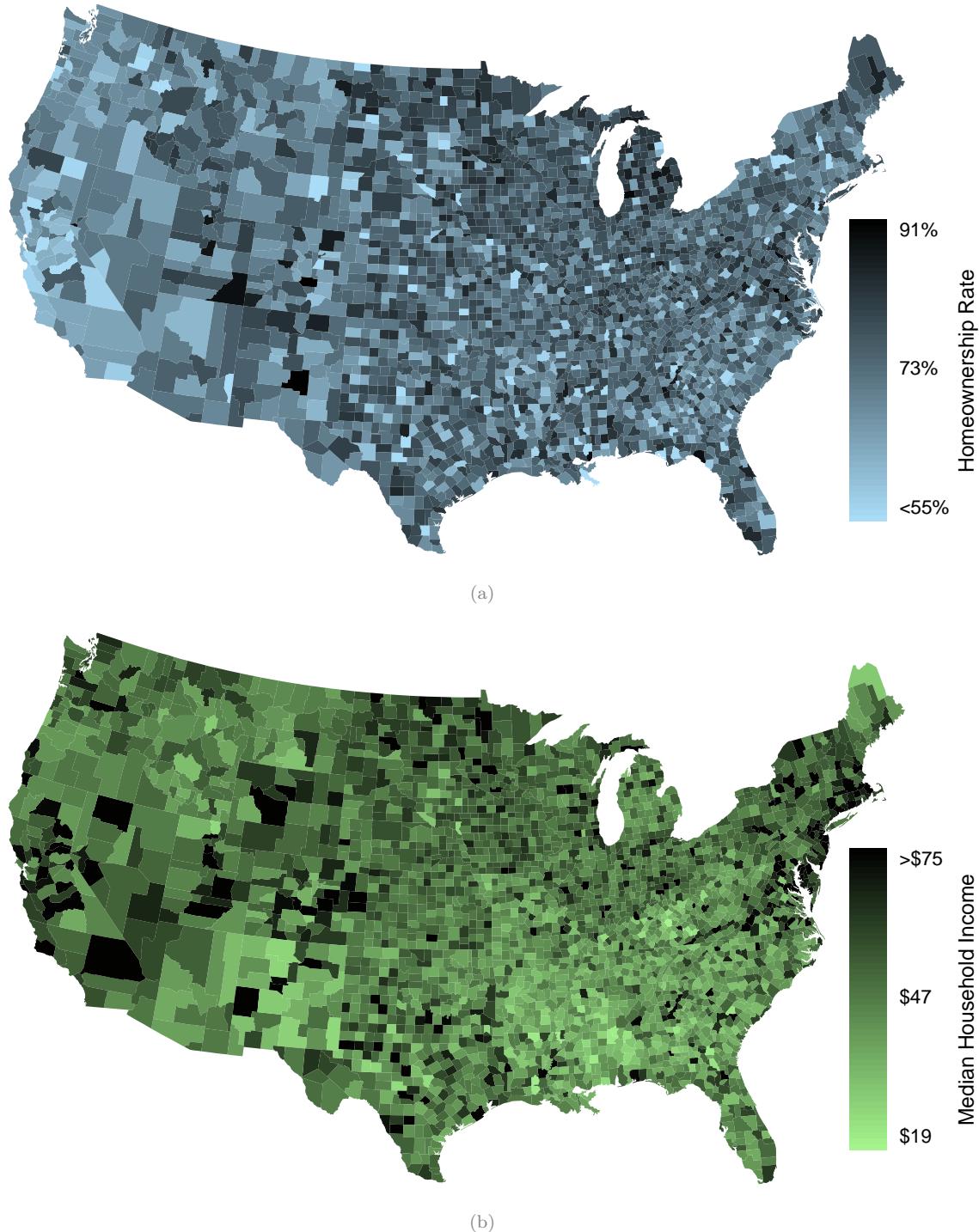
Figure 2.24: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income ($1000s). Explore dozens of intensity maps using American Community Survey data on Tableau Public ⌖.

Section summary

- In this section we looked at univariate summaries, including two measures of **center** and three measures of **spread**.

- When **summarizing** or **comparing distributions**, always comment on center, spread, and shape. Also, mention outliers or gaps if applicable. Put descriptions in *context*, that is, identify the variable(s) being summarized by name and include relevant units. Remember: *Center, Spread, and Shape! In context!*

- **Mean** and **median** are measures of center. (A common mistake is to report **mode** as a measure of center. However, a mode can appear anywhere in a distribution.)

  - The **mean** is the sum of all the observations divided by the number of observations, n.
    $$\bar{x} = \frac{1}{n} \sum x_i = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$
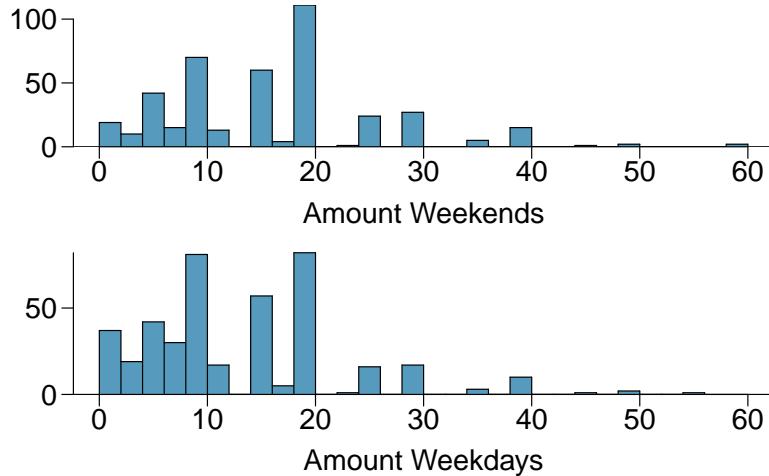
  - In an ordered data set, the **median** is the middle number when n is odd. When n is even, the median is the average of the two middle numbers.

- Because large values exert more "pull" on the mean, large values on the high end tend to increase the mean more than they increase the median. In a **right skewed** distribution, therefore, the mean is greater than the median. Analogously, in a **left skewed** distribution, the mean is less than the median. Remember: *The mean follows the tail! The skew is the tail!*

- **Standard deviation (SD)** and **Interquartile range (IQR)** are measures of spread. SD measures the typical spread from the mean, whereas IQR measures the spread of the middle 50% of the data.

  - To calculate the standard deviation, subtract the average from each value, square all those differences, add them up, divide by $n - 1$, then take the square root. Note: The standard deviation is the square root of the variance.
    $$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

  - The IQR is the difference between the third quartile Q_3 and the first quartile Q_1.
    $$IQR = Q_3 - Q_1$$

- **Range** is also sometimes used as a measure of spread. The range of a data set is defined as the difference between the maximum value and the minimum value, i.e. $max - min$.

- **Outliers** are observations that are extreme relative to the rest of the data. Two rules of thumb for identifying observations as outliers are:
  - more than 2 standard deviations above or below the mean
  - more than $1.5 \times IQR$ below Q_1 or above Q_3

  Note: These rules of thumb generally produce different cutoffs.

- Mean and SD are sensitive to outliers. Median and IQR are more robust and less sensitive to outliers.

- The **empirical rule** states that for normal distributions, about 68% of the data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean.

- **Linear transformations of data**. Adding a constant to every value in a data set shifts the mean but does not affect the standard deviation. Multiplying the values in a data set by a constant will multiply the mean and the standard deviation by that constant, except that the standard deviation must always remain positive.

- **Box plots** do not show the *distribution* of a data set in the way that histograms do. Rather, they provide a visual depiction of the **5-number summary**, which consists of: min, Q_1, Q_2, Q_3, max. It is important to be able to identify the median, IQR, and direction of skew from a box plot.

Exercises

2.7   Smoking habits of UK residents, Part I. A survey was conducted to study the smoking habits of UK residents. The histograms below display the distributions of the number of cigarettes smoked on weekdays and weekends, and they exclude data from people who identified themselves as non-smokers. Describe the two distributions and compare them.[38]

2.8   Stats scores, Part I. Below are the final exam scores of twenty introductory statistics students.

79, 83, 57, 82, 94, 83, 72, 74, 73, 71, 66, 89, 78, 81, 78, 81, 88, 69, 77, 79

Draw a histogram of these data and describe the distribution.

2.9   Smoking habits of UK residents, Part II. A random sample of 5 smokers from the data set discussed in Exercise 2.7 is provided below.

gender	age	maritalStatus	grossIncome	smoke	amtWeekends	amtWeekdays
Female	51	Married	£2,600 to £5,200	Yes	20 cig/day	20 cig/day
Male	24	Single	£10,400 to £15,600	Yes	20 cig/day	15 cig/day
Female	33	Married	£10,400 to £15,600	Yes	20 cig/day	10 cig/day
Female	17	Single	£5,200 to £10,400	Yes	20 cig/day	15 cig/day
Female	76	Widowed	£5,200 to £10,400	Yes	20 cig/day	20 cig/day

(a) Find the mean amount of cigarettes smoked on weekdays and weekends by these 5 respondents.
(b) Find the standard deviation of the amount of cigarettes smoked on weekdays and on weekends by these 5 respondents. Is the variability higher on weekends or on weekdays?

2.10   Factory defective rate. A factory quality control manager decides to investigate the percentage of defective items produced each day. Within a given work week (Monday through Friday) the percentage of defective items produced was 2%, 1.4%, 4%, 3%, 2.2%.

(a) Calculate the mean for these data.
(b) Calculate the standard deviation for these data, showing each step in detail.

2.11   Days off at a mining plant. Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

---

[38]National STEM Centre, Large Datasets from stats4schools.
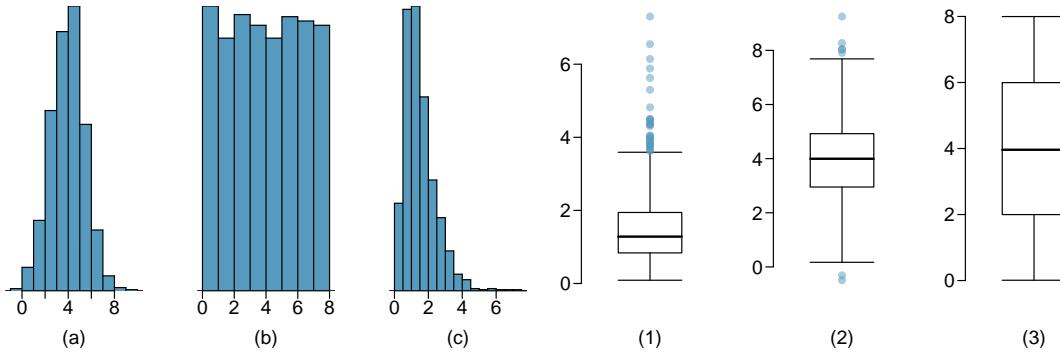
2.12   Medians and IQRs. For each part, compare distributions (1) and (2) based on their medians and
IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make
sure to explain your reasoning.

(a) (1) 3, 5, 6, 7, 9
     (2) 3, 5, 6, 7, 20
(b) (1) 3, 5, 6, 7, 9
     (2) 3, 5, 7, 8, 9

(c) (1) 1, 2, 3, 4, 5
     (2) 6, 7, 8, 9, 10
(d) (1) 0, 10, 50, 60, 100
     (2) 0, 100, 500, 600, 1000

2.13   Means and SDs. For each part, compare distributions (1) and (2) based on their means and standard
deviations. You do not need to calculate these statistics; simply state how the means and the standard
deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the
distributions.
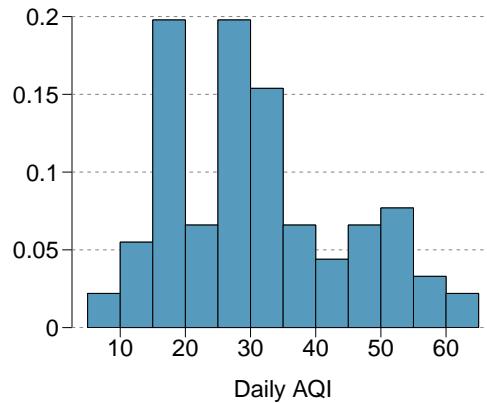
(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
     (2) 3, 5, 5, 5, 8, 11, 11, 11, 20

(b) (1) -20, 0, 0, 0, 15, 25, 30, 30
     (2) -40, 0, 0, 0, 15, 25, 30, 30

(c) (1) 0, 2, 4, 6, 8, 10
     (2) 20, 22, 24, 26, 28, 30

(d) (1) 100, 200, 300, 400, 500
     (2) 0, 50, 300, 550, 600

2.14   Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.
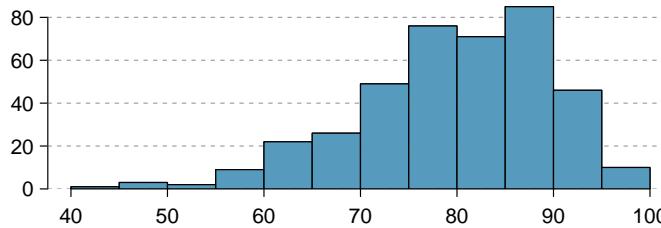
2.15   Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environ-
mental Protection Agency. This index reports the pollution level and what associated health effects might
be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes
values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of
91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI
values on these days.[39]

(a) Estimate the median AQI value of this sample.
(b) Would you expect the mean AQI value of this sample
    to be higher or lower than the median? Explain your
    reasoning.
(c) Estimate Q1, Q3, and IQR for the distribution.
(d) Would any of the days in this sample be considered
    to have an unusually low or high AQI? Explain your
    reasoning.

_____
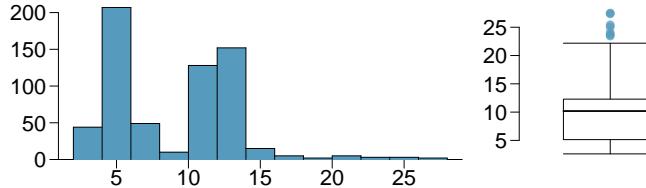
[39]US Environmental Protection Agency, AirData, 2011.

2.16   Median vs. mean. Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the median.

2.17   Histograms vs. box plots. Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?

2.18   Facebook friends. Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users?[40]

2.19   Distributions and appropriate statistics, Part I. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.
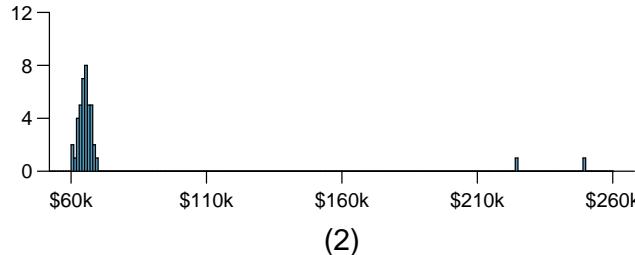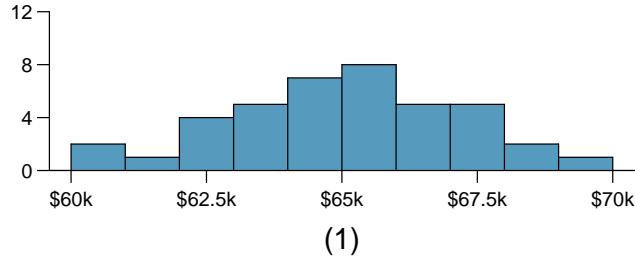
(a) Number of pets per household.

(b) Distance to work, i.e. number of miles between work and home.

(c) Heights of adult males.

2.20   Distributions and appropriate statistics, Part II. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

[40]Lars Backstrom. "Anatomy of Facebook". In: *Facebook Data Team's Notes* (2011).
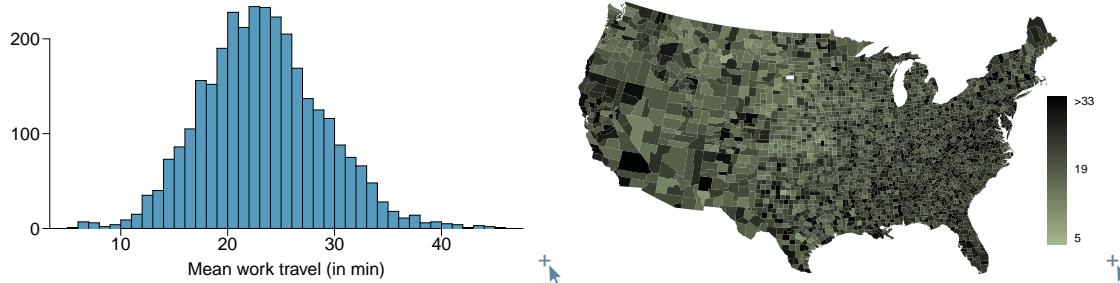
2.21 Income at the coffee shop. The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided.

(1)

(2)

	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

(a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

(b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?
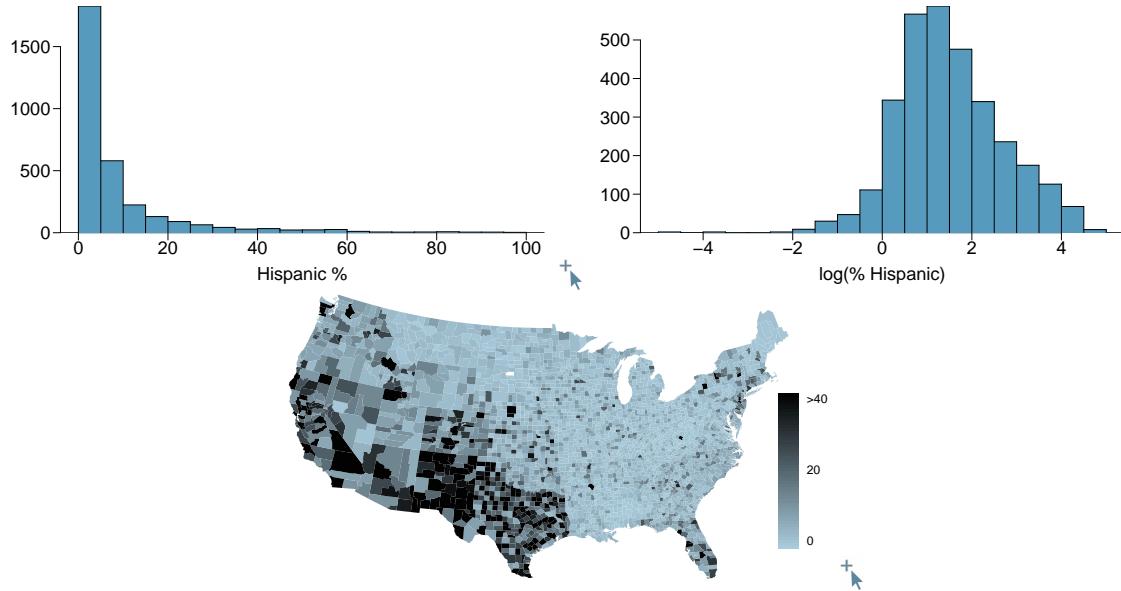
2.22 Midrange. The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning

2.23 Commute times. The US census collects data on time it takes Americans to commute to work, among many other variables. The histogram below shows the distribution of average commute times in 3,142 US counties in 2017. Also shown below is a spatial intensity map of the same data.

(a) Describe the numerical distribution and comment on whether or not a log transformation may be advisable for these data.

(b) Describe the spatial distribution of commuting times using the map provided.

2.24 Hispanic population. The US census collects data on race and ethnicity of Americans, among many other variables. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,142 counties in the US in 2017. Also shown is a histogram of logs of these values.

(a) Describe the numerical distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.

(b) What features of the distribution of the Hispanic population in US counties are apparent in the map but not in the histogram? What features are apparent in the histogram but not the map?

(c) Is one visualization more appropriate or helpful than the other? Explain your reasoning.

2.3   Considering categorical data

How do we visualize and summarize categorical data? In this section, we will introduce tables and other basic tools for categorical data that are used throughout this book and will answer the following questions:

- Based on the `loan50` data, is there an assocation between the categorical variables of homeownership and application type (individual, joint)?

- Using the `email50` data, does email type provide any useful value in classifying email as spam or not spam?

Learning objectives

1. Use a one-way table and a bar chart to summarize a categorical variable. Use counts (frequency) or proportions (relative frequency).

2. Compare distributions of a categorical variable using a two-way table and a side-by-side bar chart, segmented bar chart, or mosaic plot.

3. Calculate marginal and joint frequencies for two-way tables.

2.3.1   Contingency tables and bar charts

Figure 2.25 summarizes two variables: `app_type` and `homeownership`. A table that summarizes data for two categorical variables in this way is called a **contingency table**. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 3496 corresponds to the number of loans in the data set where the borrower rents their home and the application type was by an individual. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $3496 + 3839 + 1170 = 8505$), and **column totals** are total counts down each column. We can also create a table that shows only the overall percentages or proportions for each combination of categories, or we can create a table for a single variable, such as the one shown in Figure 2.26 for the `homeownership` variable.

		homeownership			
		rent	mortgage	own	Total
`app_type`	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Figure 2.25: A contingency table for `app_type` and `homeownership`.

A **bar chart** (also called bar plot or bar graph) is a common way to display a single categorical variable. The left panel of Figure 2.27 shows a bar chart for the `homeownership` variable. In the right panel, the counts are converted into proportions, showing the proportion of observations that are in each level (e.g. $3858/10000 = 0.3858$ for `rent`).

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Figure 2.26: A table summarizing the frequencies of each value for the `homeownership` variable.
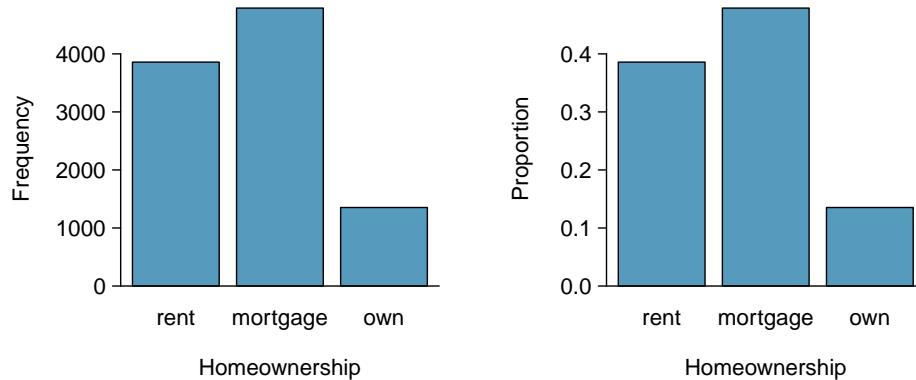
Figure 2.27: Two bar charts of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

2.3.2 Row and column proportions

Sometimes it is useful to understand the fractional breakdown of one variable in another, and we can modify our contingency table to provide such a view. Figure 2.28 shows the **row proportions** for Figure 2.25, which are computed as the counts divided by their row totals. The value 3496 at the intersection of `individual` and `rent` is replaced by $3496/8505 = 0.411$, i.e. 3496 divided by its row total, 8505. So what does 0.411 represent? It corresponds to the proportion of individual applicants who rent.

	rent	mortgage	own	Total
individual	0.411	0.451	0.138	1.000
joint	0.242	0.635	0.122	1.000
Total	0.386	0.479	0.135	1.000

Figure 2.28: A contingency table with row proportions for the `app_type` and `homeownership` variables. The row total is off by 0.001 for the `joint` row due to a rounding error.

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Figure 2.29 shows such a table, and here the value 0.906 indicates that 90.6% of renters applied as individuals for the loan. This rate is higher compared to loans from people with mortgages (80.2%) or who own their home (85.1%). Because these rates vary between the three levels of `homeownership` (`rent`, `mortgage`, `own`), this provides evidence that the `app_type` and `homeownership` variables are associated.

We could also have checked for an association between `app_type` and `homeownership` in Figure 2.28 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of loans where the borrower rents, has a mortgage, or owns varied across the `individual` to `joint` application types.

	rent	mortgage	own	Total
individual	0.906	0.802	0.865	0.851
joint	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

Figure 2.29: A contingency table with column proportions for the `app_type` and `homeownership` variables. The total for the last column is off by 0.001 due to a rounding error.

GUIDED PRACTICE 2.57

(a) What does 0.451 represent in Figure 2.28?

(b) What does 0.802 represent in Figure 2.29?[41]

GUIDED PRACTICE 2.58

(a) What does 0.122 at the intersection of `joint` and `own` represent in Figure 2.28?

(b) What does 0.135 represent in the Figure 2.29?[42]

EXAMPLE 2.59

Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One such characteristic is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is the email format, which indicates whether or not an email has any HTML content, such as bolded text. We'll focus on email format and spam status using the `email` data set, and these variables are summarized in a contingency table in Figure 2.30. Which would be more helpful to someone hoping to classify email as spam or regular email for this table: row or column proportions?

———

A data scientist would be interested in how the proportion of spam changes within each email format. This corresponds to column proportions: the proportion of spam in plain text emails and the proportion of spam in HTML emails.

If we generate the column proportions, we can see that a higher fraction of plain text emails are spam ($209/1195 = 17.5\%$) than compared to HTML emails ($158/2726 = 5.8\%$). This information on its own is insufficient to classify an email as spam or not spam, as over 80% of plain text emails are not spam. Yet, when we carefully combine this information with many other characteristics, we stand a reasonable chance of being able to classify some emails as spam or not spam with confidence.

	text	HTML	Total
spam	209	158	367
not spam	986	2568	3554
Total	1195	2726	3921

Figure 2.30: A contingency table for `spam` and `format`.

Example 2.59 points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed. However, sometimes it simply isn't clear which, if either, is more useful.

---

[41](a) 0.451 represents the proportion of individual applicants who have a mortgage. (b) 0.802 represents the fraction of applicants with mortgages who applied as individuals.

[42](a) 0.122 represents the fraction of joint borrowers who own their home. (b) 0.135 represents the home-owning borrowers who had a joint application for the loan.

EXAMPLE 2.60

Look back to Tables 2.28 and 2.29. Are there any obvious scenarios where one might be more useful than the other?

———————

None that we thought were obvious! What is distinct about `app_type` and `homeownership` vs the email example is that these two variables don't have a clear explanatory-response variable relationship that we might hypothesize (see Section 1.3.3 for these terms). Usually it is most useful to "condition" on the explanatory variable. For instance, in the email example, the email format was seen as a possible explanatory variable of whether the message was spam, so we would find it more interesting to compute the relative frequencies (proportions) for each email format.

2.3.3 Using a bar chart with two variables

Contingency tables using row or column proportions are especially useful for examining how two categorical variables are related. Segmented bar charts provide a way to visualize the information in these tables.

A **segmented bar chart**, or stacked bar chart, is a graphical display of contingency table information. For example, a segmented bar chart representing Figure 2.29 is shown in Figure 2.31(a), where we have first created a bar chart using the `homeownership` variable and then divided each group by the levels of `app_type`.

One related visualization to the segmented bar chart is the **side-by-side bar chart**, where an example is shown in Figure 2.31(b).

For the last type of bar chart we introduce, the column proportions for the `app_type` and `homeownership` contingency table have been translated into a standardized segmented bar chart in Figure 2.31(c). This type of visualization is helpful in understanding the fraction of individual or joint loan applications for borrowers in each level of `homeownership`. Additionally, since the proportions of `joint` and `individual` vary across the groups, we can conclude that the two variables are associated.

EXAMPLE 2.61

Examine the four bar charts in Figure 2.31. When is the segmented, side-by-side, standardized segmented bar chart, or standardized side-by-side the most useful?

———————

The segmented bar chart is most useful when it's reasonable to assign one variable as the explanatory variable and the other variable as the response, since we are effectively grouping by one variable first and then breaking it down by the others.

Side-by-side bar charts are more agnostic in their display about which variable, if any, represents the explanatory and which the response variable. It is also easy to discern the number of cases in of the six different group combinations. However, one downside is that it tends to require more horizontal space; the narrowness of Figure 2.31(b) makes the plot feel a bit cramped. Additionally, when two groups are of very different sizes, as we see in the `own` group relative to either of the other two groups, it is difficult to discern if there is an association between the variables.

The standardized segmented bar chart is helpful if the primary variable in the segmented bar chart is relatively imbalanced, e.g. the `own` category has only a third of the observations in the `mortgage` category, making the simple segmented bar chart less useful for checking for an association. The major downside of the standardized version is that we lose all sense of how many cases each of the bars represents.

The last plot is a standardized side-by-side bar chart. It shows the joint and individual groups as proportions within each level of homeownership, and it offers similar benefits and tradeoffs to the standardized version of the stacked bar plot.

(a)

(b)

(c)

(d)
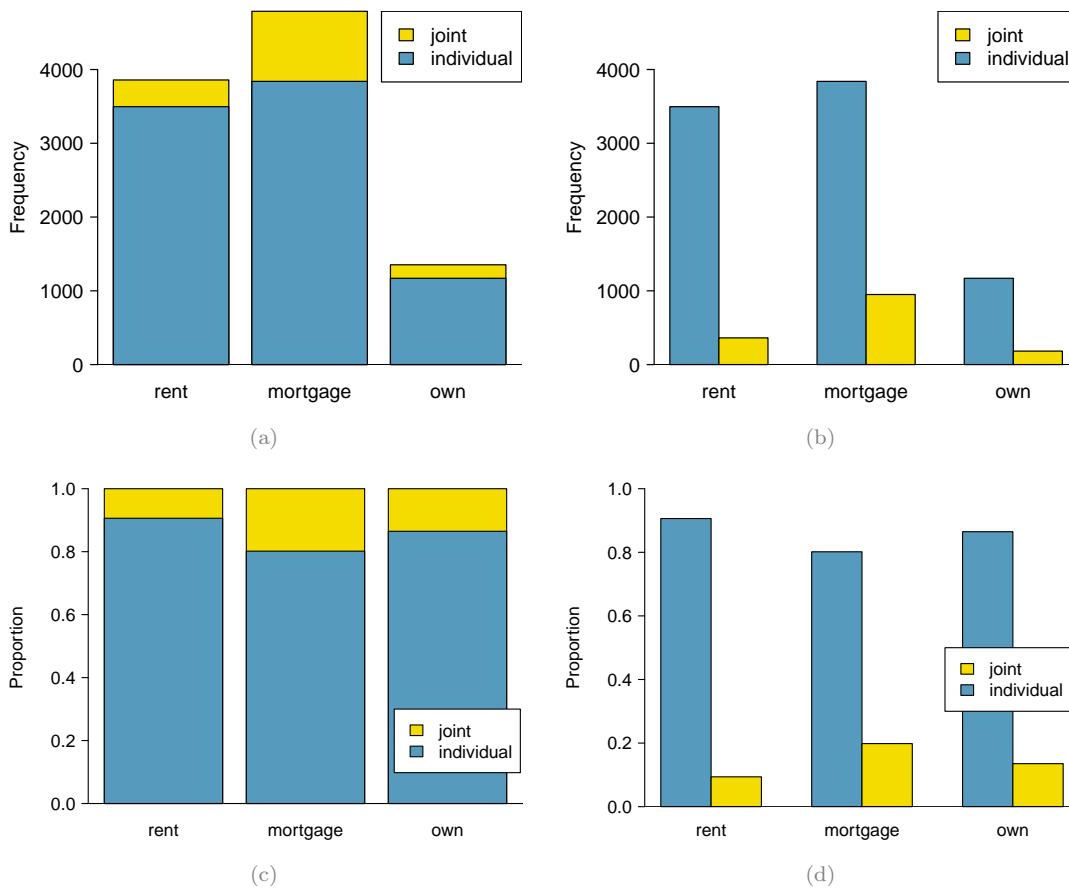
Figure 2.31: (a) segmented bar chart for `homeownership`, where the counts have been further broken down by `app_type`. (b) Side-by-side bar chart for `homeownership` and `app_type`. (c) Standardized version of the segmented bar chart. (d) Standardized side-by-side bar chart. See these bar charts on Tableau Public ⁺↖.

2.3.4 Mosaic plots

A **mosaic plot** is a visualization technique suitable for contingency tables that resembles a standardized segmented bar chart with the benefit that we still see the relative group sizes of the primary variable as well.

To get started in creating our first mosaic plot, we'll break a square into columns for each category of the `homeownership` variable, with the result shown in Figure 2.32(a). Each column represents a level of `homeownership`, and the column widths correspond to the proportion of loans in each of those categories. For instance, there are fewer loans where the borrower is an owner than where the borrower has a mortgage. In general, mosaic plots use box *areas* to represent the number of cases in each category.

To create a completed mosaic plot, the single-variable mosaic plot is further divided into pieces in Figure 2.32(b) using the `app_type` variable. Each column is split proportional to the number of loans from individual and joint borrowers. For example, the second column represents loans where the borrower has a mortgage, and it was divided into individual loans (upper) and joint loans (lower). As another example, the bottom segment of the third column represents loans where the borrower owns their home and applied jointly, while the upper segment of this column represents borrowers who are homeowners and filed individually. We can again use this plot to see that the `homeownership` and `app_type` variables are associated, since some columns are divided in different vertical locations than others, which was the same technique used for checking an association in the standardized segmented bar chart.

Figure 2.32: (a) The one-variable mosaic plot for `homeownership`. (b) Two-variable mosaic plot for both `homeownership` and `app_type`.

In Figure 2.33, we chose to first split by the homeowner status of the borrower. However, we could have instead first split by the application type, as in Figure 2.33. Like with the bar charts, it's common to use the explanatory variable to represent the first split in a mosaic plot, and then for the response to break up each level of the explanatory variable, if these labels are reasonable to attach to the variables under consideration.

Figure 2.33: Mosaic plot where loans are grouped by the `homeownership` variable after they've been divided into the `individual` and `joint` application types.

2.3.5   The only pie chart you will see in this book

A **pie chart** is shown in Figure 2.34 alongside a bar chart representing the same information. Pie charts can be useful for giving a high-level overview to show how a set of cases break down. However, it is also difficult to decipher details in a pie chart. For example, it takes a couple seconds longer to recognize that there are more loans where the borrower has a mortgage than rent when looking at the pie chart, while this detail is very obvious in the bar chart. While pie charts can be useful, we prefer bar charts for their ease in comparing groups.
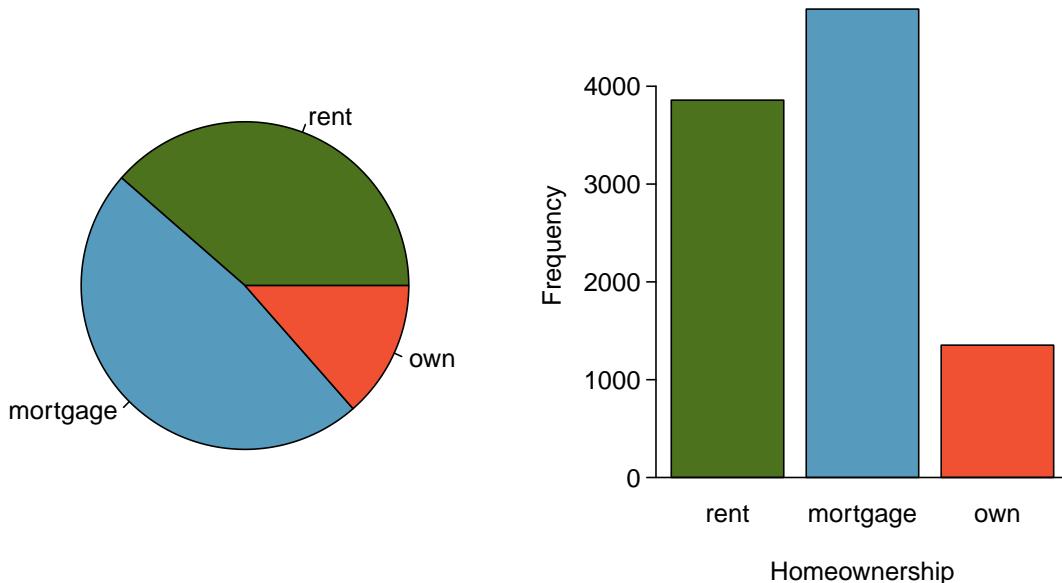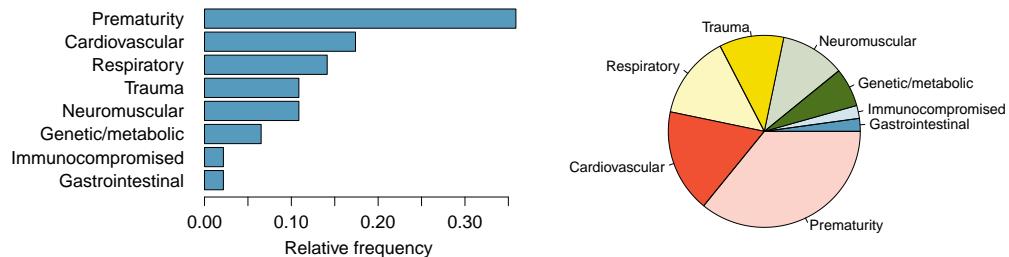
Figure 2.34: A pie chart and bar chart of `homeownership`. Compare multiple ways of summarizing a single categorical variable on Tableau Public .

Section summary

- **Categorical variables**, unlike numerical variables, are simply summarized by **counts** (how many) and **proportions**. These are referred to as frequency and relative frequency, respectively.

- When summarizing one categorical variable, a **one-way frequency table** is useful. For summarizing two categorical variables and their relationship, use a **two-way frequency table** (also known as a contingency table).

- To graphically summarize a single categorical variable, use a **bar chart**. To summarize and compare two categorical variables, use a **side-by-side bar chart**, a **segmented bar chart**, or a **mosaic plot**.

- **Pie charts** are another option for summarizing categorical data, but they are more difficult to read and bar charts are generally a better option.

Exercises

2.25 Antibiotic use in children. The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.

(a) What features are apparent in the bar plot but not in the pie chart?

(b) What features are apparent in the pie chart but not in the bar plot?

(c) Which graph would you prefer to use for displaying these categorical data?
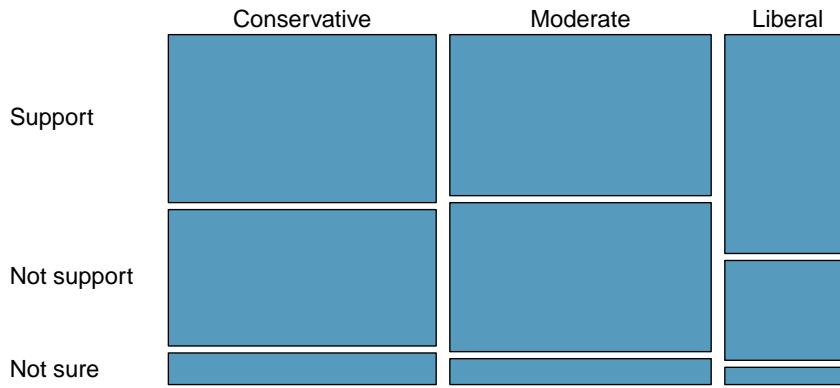
2.26 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[43]

		Political ideology		
		Conservative	Moderate	Liberal	Total
	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
Response	(iii) Leave the country	179	126	45	350
	(iv) Not sure	15	4	1	20
	Total	372	363	175	910
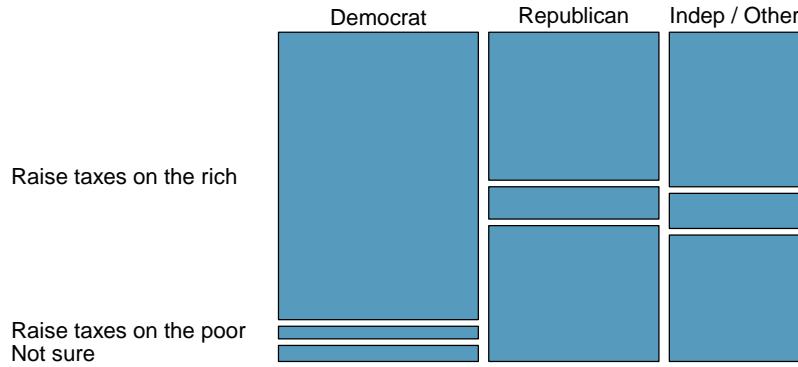
(a) What percent of these Tampa, FL voters identify themselves as conservatives?

(b) What percent of these Tampa, FL voters are in favor of the citizenship option?

(c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

(d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

---

[43]SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

2.27   Views on the DREAM Act. A random sample of registered voters from Tampa, FL were asked if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. The survey also collected information on the political ideology of the respondents. Based on the mosaic plot shown below, do views on the DREAM Act and political ideology appear to be independent? Explain your reasoning.[44]

2.28   Raise taxes. A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. The survey also collected information on the political party affiliation of the respondents. Based on the mosaic plot shown below, do views on raising taxes and political affiliation appear to be independent? Explain your reasoning.[45]

---

[44]SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.
[45]Public Policy Polling, Americans on College Degrees, Classic Literature, the Seasons, and More, data collected Feb 20-22, 2015.

2.4 Case study: malaria vaccine (special topic)

How large does an observed difference need to be for it to provide convincing evidence that something real is going on, something beyond random variation? Answering this question requires the tools that we will encounter in the later chapters on probability and inference. However, this is such an interesting and important question, and we'll also address it here using simulation. This section can be covered now or in tandem with Chapter 5: Foundations for Inference.

Learning objectives

1. Recognize that an observed difference in sample statistics may be due to random chance and that we use hypothesis testing to determine if this difference statistically significant (i.e. too large to be attributed to random chance).

2. Set up competing hypotheses and use the results of a simulation to evaluate the degree of support the data provide against the null hypothesis and for the alternative hypothesis.

2.4.1 Variability within data

EXAMPLE 2.62

Suppose your professor splits the students in class into two groups: students on the left and students on the right. If $\hat{p}_L$ and $\hat{p}_R$ represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if $\hat{p}_L$ did not exactly equal $\hat{p}_R$?

———

While the proportions would probably be close to each other, it would be unusual for them to be exactly the same. We would probably observe a small difference due to chance.

GUIDED PRACTICE 2.63

If we don't think the side of the room a person sits on in class is related to whether the person owns an Apple product, what assumption are we making about the relationship between these two variables?[46]

We consider a study on a new malaria vaccine called PfSPZ. In this study, volunteer patients were randomized into one of two experiment groups: 14 patients received an experimental vaccine or 6 patients received a placebo vaccine. Nineteen weeks later, all 20 patients were exposed to a drug-sensitive malaria virus strain; the motivation of using a drug-sensitive strain of virus here is for ethical considerations, allowing any infections to be treated effectively. The results are summarized in Figure 2.35, where 9 of the 14 treatment patients remained free of signs of infection while all of the 6 patients in the control group patients showed some baseline signs of infection.

———

[46]We would be assuming that these two variables are independent.

| | outcome | | |
	infection	no infection	Total
vaccine	5	9	14
placebo	6	0	6
Total	11	9	20

Figure 2.35: Summary results for the malaria vaccine experiment.

GUIDED PRACTICE 2.64

Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[47]

In this study, a smaller proportion of patients who received the vaccine showed signs of an infection (35.7% versus 100%). However, the sample is very small, and it is unclear whether the difference provides *convincing evidence* that the vaccine is effective.

EXAMPLE 2.65

Data scientists are sometimes called upon to evaluate the strength of evidence. When looking at the rates of infection for patients in the two groups in this study, what comes to mind as we try to determine whether the data show convincing evidence of a real difference?

The observed infection rates (35.7% for the treatment group versus 100% for the control group) suggest the vaccine may be effective. However, we cannot be sure if the observed difference represents the vaccine's efficacy or is just from random chance. Generally there is a little bit of fluctuation in sample data, and we wouldn't expect the sample proportions to be *exactly* equal, even if the truth was that the infection rates were independent of getting the vaccine. Additionally, with such small samples, perhaps it's common to observe such large differences when we randomly split a group due to chance alone!

Example 2.65 is a reminder that the observed outcomes in the data sample may not perfectly reflect the true relationships between variables since there is **random noise**. While the observed difference in rates of infection is large, the sample size for the study is small, making it unclear if this observed difference represents efficacy of the vaccine or whether it is simply due to chance. We label these two competing claims, H_0 and H_A, which are spoken as "H-nought" and "H-A":

H_0: **Independence model.** The variables treatment and outcome are independent. They have no relationship, and the observed difference between the proportion of patients who developed an infection in the two groups, 64.3%, was due to chance.

H_A: **Alternative model.** The variables are *not* independent. The difference in infection rates of 64.3% was not due to chance, and vaccine affected the rate of infection.

What would it mean if the independence model, which says the vaccine had no influence on the rate of infection, is true? It would mean 11 patients were going to develop an infection *no matter which group they were randomized into*, and 9 patients would not develop an infection *no matter which group they were randomized into*. That is, if the vaccine did not affect the rate of infection, the difference in the infection rates was due to chance alone in how the patients were randomized.

Now consider the alternative model: infection rates were influenced by whether a patient received the vaccine or not. If this was true, and especially if this influence was substantial, we would expect to see some difference in the infection rates of patients in the groups.

We choose between these two competing claims by assessing if the data conflict so much with H_0 that the independence model cannot be deemed reasonable. If this is the case, and the data support H_A, then we will reject the notion of independence and conclude there was discrimination.

---

[47]The study is an experiment, as patients were randomly assigned an experiment group. Since this is an experiment, the results can be used to evaluate a causal relationship between the malaria vaccine and whether patients showed signs of an infection.

2.4.2 Simulating the study

We're going to implement **simulations**, where we will pretend we know that the malaria vaccine being tested does *not* work. Ultimately, we want to understand if the large difference we observed is common in these simulations. If it is common, then maybe the difference we observed was purely due to chance. If it is very uncommon, then the possibility that the vaccine was helpful seems more plausible.

Figure 2.35 shows that 11 patients developed infections and 9 did not. For our simulation, we will suppose the infections were independent of the vaccine and we were able to *rewind* back to when the researchers randomized the patients in the study. If we happened to randomize the patients differently, we may get a different result in this hypothetical world where the vaccine doesn't influence the infection. Let's complete another **randomization** using a simulation.

In this **simulation**, we take 20 notecards to represent the 20 patients, where we write down "infection" on 11 cards and "no infection" on 9 cards. In this hypothetical world, we believe each patient that got an infection was going to get it regardless of which group they were in, so let's see what happens if we randomly assign the patients to the treatment and control groups again. We thoroughly shuffle the notecards and deal 14 into a `vaccine` pile and 6 into a `placebo` pile. Finally, we tabulate the results, which are shown in Figure 2.36.

		outcome		
		infection	no infection	Total
treatment	vaccine	7	7	14
(simulated)	placebo	4	2	6
	Total	11	9	20

Figure 2.36: Simulation results, where any difference in infection rates is purely due to chance.

GUIDED PRACTICE 2.66

Ⓖ What is the difference in infection rates between the two simulated groups in Figure 2.36? How does this compare to the observed 64.3% difference in the actual data?[48]

2.4.3 Checking for independence

We computed one possible difference under the independence model in Guided Practice 2.66, which represents one difference due to chance. While in this first simulation, we physically dealt out notecards to represent the patients, it is more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance:

$$\frac{2}{6} - \frac{9}{14} = -0.310$$

And another:

$$\frac{3}{6} - \frac{8}{14} = -0.071$$

And so on until we repeat the simulation enough times that we have a good idea of what represents the *distribution of differences from chance alone*. Figure 2.37 shows a stacked plot of the differences found from 100 simulations, where each dot represents a simulated difference between the infection rates (control rate minus treatment rate).

Note that the distribution of these simulated differences is centered around 0. We simulated these differences assuming that the independence model was true, and under this condition, we

---

[48] $4/6 - 7/14 = 0.167$ or about 16.7% in favor of the vaccine. This difference due to chance is much smaller than the difference observed in the actual groups.
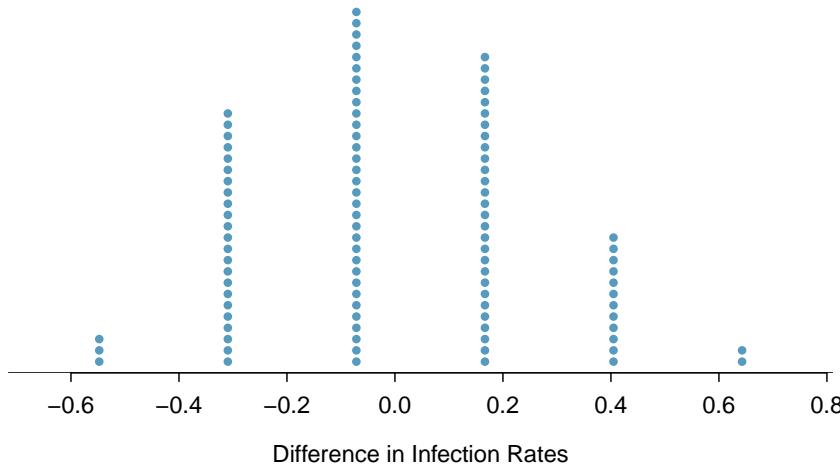
Figure 2.37: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

expect the difference to be near zero with some random fluctuation, where *near* is pretty generous in this case since the sample sizes are so small in this study.

EXAMPLE 2.67

Given the results of the simulation shown in Figure 2.37, about how often would you expect to observe a result as large as 64.3% if H_0 were true?

Because a result this large happened 2 times out the 100 simulations, we would expect such a large value only 2% of the time if H_0 were true.

There are two possible interpretations of the results of the study:

H_0 **Independence model.** The vaccine has no effect on infection rate, and we just happened to observe a rare event.

H_A **Alternative model.** The vaccine has an effect on infection rate, and the difference we observed was actually due to the vaccine being effective at combatting malaria, which explains the large difference of 64.3%.

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong enough evidence against the independence model, meaning we do not conclude that the vaccine had an effect in this clinical setting. (2) We conclude the evidence is sufficiently strong to reject H_0, and we assert that the vaccine was useful.

Is 2% small enough to make us reject the independence model? That depends on how much evidence we require. The smaller that probability is, the more evidence it provides against H_0. Later, we will see that researchers often use a cutoff of 5%, though it can depend upon the situation. Using the 5% cutoff, we would reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence that the vaccine provides some protection against malaria in this clinical setting.

When there is strong enough evidence that the result points to a real difference and is not simply due to random variation, we call the result **statistically significant**.

One field of statistics, statistical inference, is built on evaluating whether such differences are due to chance. In statistical inference, data scientists evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur. In Chapter 5, we give a formal introduction to the problem of model selection. We spend the next two chapters building a foundation of probability and theory necessary to make that discussion rigorous.
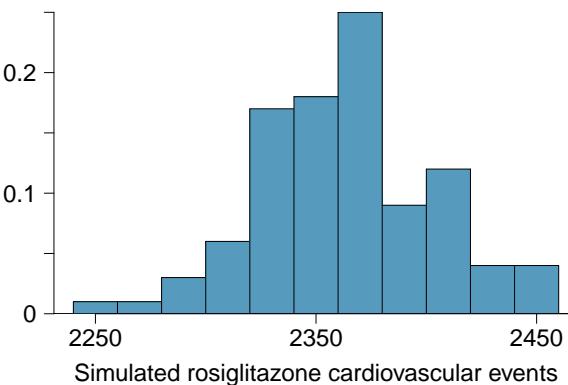
Exercises

2.29 Side effects of Avandia. Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.[49]

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571
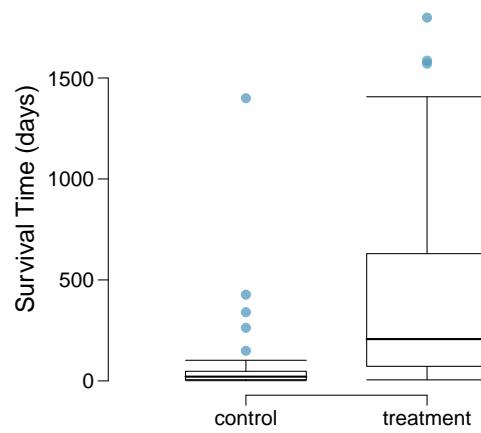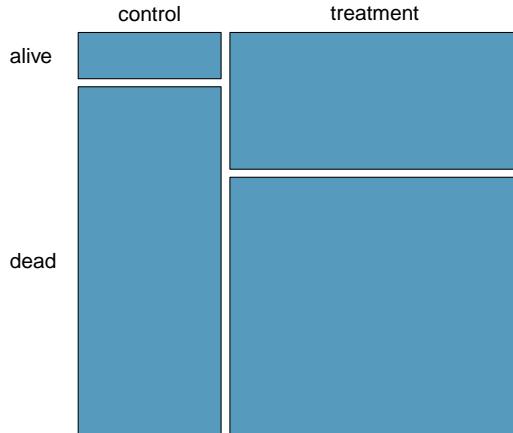
(a) Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

   i. Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

   ii. The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

   iii. The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

   iv. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

(b) What proportion of all patients had cardiovascular problems?

(c) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

(d) We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 1,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Use the relative frequency histogram of these counts to answer (i)-(iii).

   i. What are the claims being tested?

   ii. Compared to the number calculated in part (b), which would provide more support for the alternative hypothesis, *more* or *fewer* patients with cardiovascular problems in the rosiglitazone group?

   iii. What do the simulation results suggest about the relationship between taking rosiglitazone and having cardiovascular problems in diabetic patients?

Simulated rosiglitazone cardiovascular events

---

[49]D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: *JAMA* 304.4 (2010), p. 411. ISSN: 0098-7484.
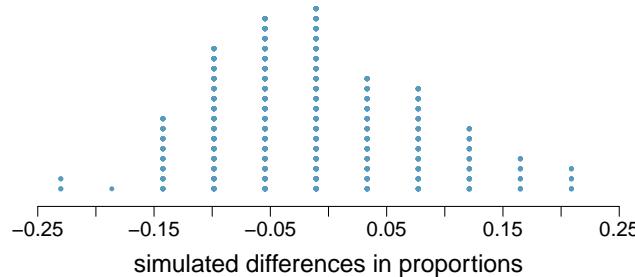
2.30   Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study.[50]

(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

  i. What are the claims being tested?

  ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

    We write *alive* on _____ cards representing patients who were alive at the end of the study, and *dead* on _____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size _____ representing treatment, and another group of size _____ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at _____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

  iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

simulated differences in proportions

---

[50]B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

Chapter highlights

A raw data matrix/table may have thousands of rows. The data need to be summarized in order to makes sense of all the information. In this chapter, we looked at ways to summarize data **graphically**, **numerically**, and **verbally**.

Categorical data

- A single **categorical variable** is summarized with **counts** or **proportions** in a **one-way table**. A **bar chart** is used to show the frequency or relative frequency of the categories that the variable takes on.

- Two categorical variables can be summarized in a **two-way table** and with a **side-by-side bar chart** or a **segmented bar chart**.

Numerical data

- When looking at a single **numerical variable**, we try to understand the **distribution** of the variable. The distribution of a variable can be represented with a frequency table and with a graph, such as a **stem-and-leaf plot** or **dot plot** for small data sets, or a **histogram** for larger data sets. If only a summary is desired, a **box plot** may be used.

- The **distribution** of a variable can be described and summarized with **center** (mean or median), **spread** (SD or IQR), and **shape** (right skewed, left skewed, approximately symmetric).

- **Z-scores** and **percentiles** are useful for identifying a data point's relative position within a data set.

- **Outliers** are values that appear extreme relative to the rest of the data. Investigating outliers can provide insight into properties of the data or may reveal data collection/entry errors.

- When **comparing the distribution** of two variables, use two dot plots, two histograms, a back-to-back stem-and-leaf, or parallel box plots.

- To look at the **association** between two numerical variables, use a **scatterplot**.

Graphs and numbers can summarize data, but they alone are insufficient. It is the role of the researcher or data scientist to ask questions, to use these tools to identify patterns and departure from patterns, and to make sense of this in the context of the data. Strong writing skills are critical for being able to communicate the results to a wider audience.
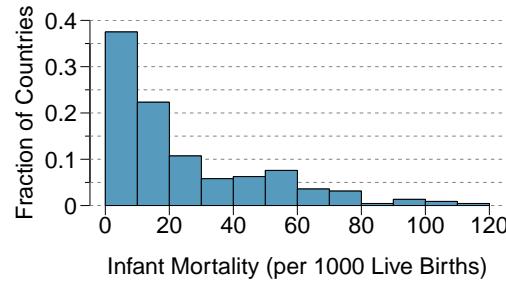
Chapter exercises

2.31 Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

(a) Does the new student's score increase or decrease the average score?
(b) What is the new average?
(c) Does the new student's score increase or decrease the standard deviation of the scores?

2.32 Infant mortality. The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.[51]

(a) Estimate Q1, the median, and Q3 from the histogram.
(b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.
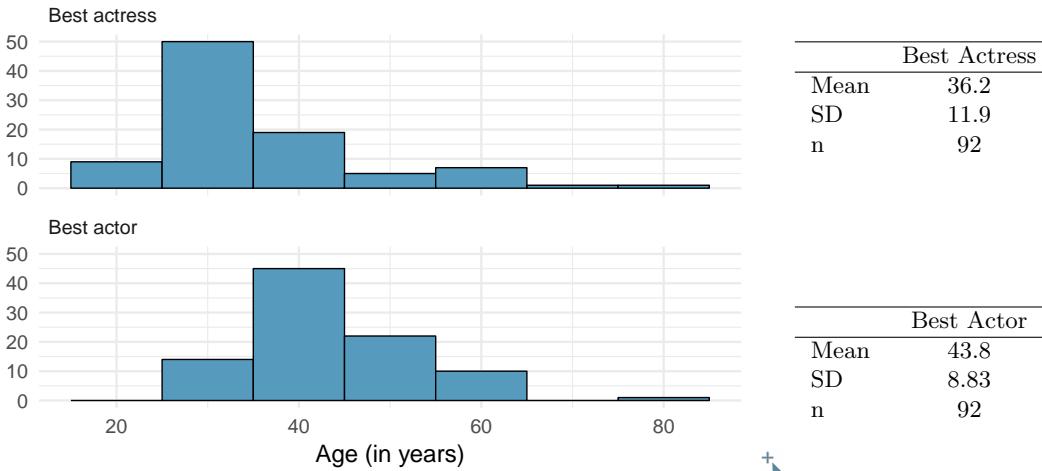
2.33 TV watchers. Students in an AP Statistics class were asked how many hours of television they watch per week (including online streaming). This sample yielded an average of 4.71 hours, with a standard deviation of 4.18 hours. Is the distribution of number of hours students watch television weekly symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

2.34 A new statistic. The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are greater than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.

(a) $\frac{\bar{x}}{median} = 1$
(b) $\frac{\bar{x}}{median} < 1$
(c) $\frac{\bar{x}}{median} > 1$

2.35 Oscar winners. The first Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2018. Summary statistics for these distributions are also provided. Compare the distributions of ages of best actor and actress winners.[52]

	Best Actress
Mean	36.2
SD	11.9
n	92

	Best Actor
Mean	43.8
SD	8.83
n	92

---

[51]CIA Factbook, Country Comparisons, 2014.
[52]Oscar winners from 1929 – 2018, data up to 2009 from the Journal of Statistics Education data archive and more current data from wikipedia.org.

2.36   Exam scores. The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15. Is the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.
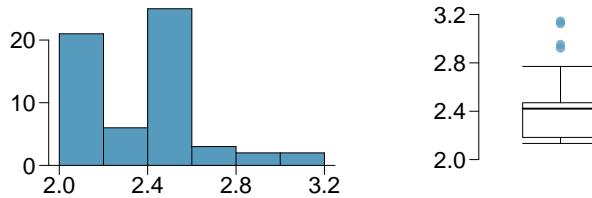
2.37   Stats scores. Below are the final exam scores of twenty introductory statistics students.

57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94
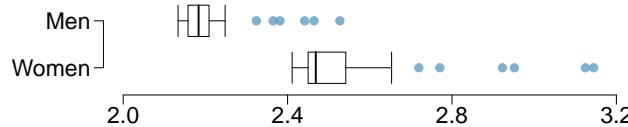
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94
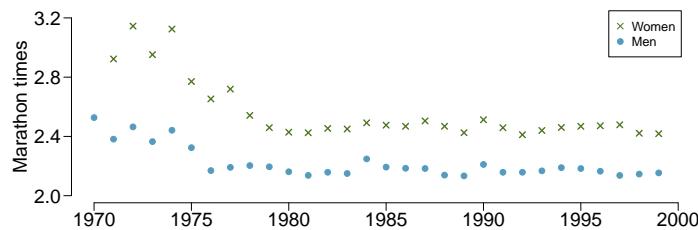
2.38   Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.

(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

(c) Compare the distribution of marathon times for men and women based on the box plot shown below.

(d) The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.

Chapter 3

Probability

Probability forms a foundation of statistics, and you're probably already aware of many of the ideas. However, formalization of the concepts is new for most. This chapter aims to introduce probability concepts through examples that will be familiar to most people.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

3.1   Defining probability

What is the probability of rolling an even number on a die? Of getting 5 heads in row when tossing a coin? Of drawing a Heart or an Ace from a deck of cards? The study of probability is fun and interesting in its own right, but it also forms the foundation for statistical models and inferential procedures, many of which we will investigate in subsequent chapters.

Learning objectives

1. Describe the long-run relative frequency interpretation of probability and understand its relationship to the "Law of Large Numbers".

2. Use Venn diagrams to represent events and their probabilities and to visualize the complement, union, and intersection of events.

3. Use the General Addition Rule to find the probability that at least one of several events occurs.

4. Understand when events are disjoint (mutually exclusive) and how that simplifies the General Addition Rule.

5. Apply the Multiplication Rule for finding the joint probability of independent events.

3.1.1   Introductory examples

EXAMPLE 3.1

A "die", the singular of dice, is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting 1 when rolling a die?

———————

If the die is fair, then the chance of a 1 is as good as the chance of any other number. Since there are six outcomes, the chance must be 1-in-6 or, equivalently, 1/6.

EXAMPLE 3.2

What is the chance of getting a 1 or 2 in the next roll?

———————

1 and 2 constitute two of the six equally likely possible outcomes, so the chance of getting one of these two outcomes must be 2/6 = 1/3.

EXAMPLE 3.3

What is the chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?

———————

100%. The outcome must be one of these numbers.

EXAMPLE 3.4

What is the chance of not rolling a 2?

Since the chance of rolling a 2 is 1/6 or $16.\bar{6}\%$, the chance of not rolling a 2 must be $100\% - 16.\bar{6}\% = 83.\bar{3}\%$ or 5/6.

Alternatively, we could have noticed that not rolling a 2 is the same as getting a 1, 3, 4, 5, or 6, which makes up five of the six equally likely outcomes and has probability 5/6.

EXAMPLE 3.5

Consider rolling two dice. If $1/6^{th}$ of the time the first die is a 1 and $1/6^{th}$ of those times the second die is a 1, what is the chance of getting two 1s?

If $16.\bar{6}\%$ of the time the first die is a 1 and $1/6^{th}$ of *those* times the second die is also a 1, then the chance that both dice are 1 is $(1/6) \times (1/6)$ or 1/36.

3.1.2 Probability

We use probability to build tools to describe and understand apparent randomness. We often frame probability in terms of a **random process** giving rise to an **outcome**.

$$\text{Roll a die} \quad \rightarrow \quad \text{1, 2, 3, 4, 5, or 6}$$
$$\text{Flip a coin} \quad \rightarrow \quad \text{H or T}$$

Rolling a die or flipping a coin is a seemingly random process and each gives rise to an outcome.

> **PROBABILITY**
>
> The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be displayed as a percentage between 0% and 100%.

Probability can be illustrated by rolling a die many times. Consider the event "roll a 1". The **relative frequency** of an event is the proportion of times the event occurs out of the number of trials. Let $\hat{p}_n$ be the proportion of outcomes that are 1 after the first n rolls. As the number of rolls increases, $\hat{p}_n$ (the relative frequency of rolls) will converge to the probability of rolling a 1, $p = 1/6$. Figure 3.1 shows this convergence for 100,000 die rolls. The tendency of $\hat{p}_n$ to stabilize around p, that is, the tendency of the relative frequency to stabilize around the true probability, is described by the **Law of Large Numbers**.

> **LAW OF LARGE NUMBERS**
>
> As more observations are collected, the observed proportion $\hat{p}_n$ of occurrences with a particular outcome after n trials converges to the true probability p of that outcome.

Occasionally the proportion will veer off from the probability and appear to defy the Law of Large Numbers, as $\hat{p}_n$ does many times in Figure 3.1. However, these deviations become smaller as the number of rolls increases.

Above we write p as the probability of rolling a 1. We can also write this probability as

$$P(\text{rolling a 1})$$

As we become more comfortable with this notation, we will abbreviate it further. For instance, if it is clear that the process is "rolling a die", we could abbreviate $P(\text{rolling a 1})$ as $P(1)$.
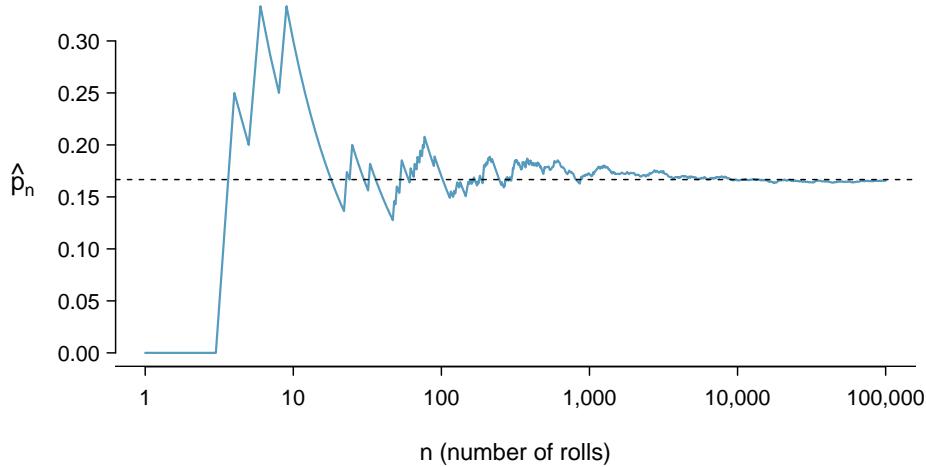
Figure 3.1: The fraction of die rolls that are 1 at each stage in a simulation. The relative frequency tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

GUIDED PRACTICE 3.6

Random processes include rolling a die and flipping a coin. (a) Think of another random process. (b) Describe all the possible outcomes of that process. For instance, rolling a die is a random process with potential outcomes 1, 2, ..., 6. [1]

What we think of as random processes are not necessarily random, but they may just be too difficult to understand exactly. The fourth example in the footnote solution to Guided Practice 3.6 suggests a roommate's behavior is a random process. However, even if a roommate's behavior is not truly random, modeling her behavior as a random process can still be useful.

MODELING A PROCESS AS RANDOM

It can be helpful to model a process as random even if it is not truly random.

3.1.3   Disjoint or mutually exclusive outcomes

Two outcomes are called **disjoint** or **mutually exclusive** if they cannot both happen in the same trial. For instance, if we roll a die, the outcomes 1 and 2 are disjoint since they cannot both occur on a single roll. On the other hand, the outcomes 1 and "rolling an odd number" are not disjoint since both occur if the outcome of the roll is a 1. The terms *disjoint* and *mutually exclusive* are equivalent and interchangeable.

Calculating the probability of disjoint outcomes is easy. When rolling a die, the outcomes 1 and 2 are disjoint, and we compute the probability that one of these outcomes will occur by adding their separate probabilities:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

What about the probability of rolling a 1, 2, 3, 4, 5, or 6? Here again, all of the outcomes are

---

[1]Here are four examples. (i) Whether someone gets sick in the next month or not is an apparently random process with outcomes sick and not. (ii) We can *generate* a random process by randomly picking a person and measuring that person's height. The outcome of this process will be a positive number. (iii) Whether the stock market goes up or down next week is a seemingly random process with possible outcomes up, down, and no_change. Alternatively, we could have used the percent change in the stock market as a numerical outcome. (iv) Whether your roommate cleans her dishes tonight probably seems like a random process with possible outcomes cleans_dishes and leaves_dishes.

disjoint so we add the probabilities:

$$P(\texttt{1 or 2 or 3 or 4 or 5 or 6})$$
$$= P(\texttt{1}) + P(\texttt{2}) + P(\texttt{3}) + P(\texttt{4}) + P(\texttt{5}) + P(\texttt{6})$$
$$= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1.$$

The **Addition Rule** guarantees the accuracy of this approach when the outcomes are disjoint.

ADDITION RULE OF DISJOINT OUTCOMES

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k) \tag{3.7}$$

GUIDED PRACTICE 3.8

We are interested in the probability of rolling a 1, 4, or 5. (a) Explain why the outcomes 1, 4, and 5 are disjoint. (b) Apply the Addition Rule for disjoint outcomes to determine $P(\texttt{1 or 4 or 5})$.[2]

GUIDED PRACTICE 3.9

In the `email` data set in Chapter 2, the `number` variable described whether no number (labeled `none`), only one or more small numbers (`small`), or whether at least one big number appeared in an email (`big`). Of the 3,921 emails, 549 had no numbers, 2,827 had only one or more small numbers, and 545 had at least one big number. (a) Are the outcomes `none`, `small`, and `big` disjoint? (b) Determine the proportion of emails with value `small` and `big` separately. (c) Use the Addition Rule for disjoint outcomes to compute the probability a randomly selected email from the data set has a number in it, small or big.[3]

Statisticians rarely work with individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where a die roll results in 1 or 2 and B represent the event that the die roll is a 4 or a 6. We write A as the set of outcomes $\{1, 2\}$ and $B = \{4, 6\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 3.2.
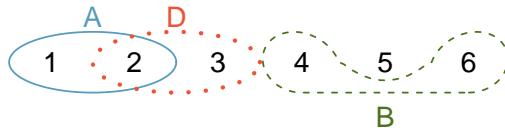
Figure 3.2: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

---

[2](a) The random process is a die roll, and at most one of these outcomes can come up. This means they are disjoint outcomes. (b) $P(\texttt{1 or 4 or 5}) = P(\texttt{1}) + P(\texttt{4}) + P(\texttt{5}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$.

[3](a) Yes. Each email is categorized in only one level of `number`. (b) Small: $\frac{2827}{3921} = 0.721$. Big: $\frac{545}{3921} = 0.139$. (c) $P(\texttt{small or big}) = P(\texttt{small}) + P(\texttt{big}) = 0.721 + 0.139 = 0.860$.

The Addition Rule applies to both disjoint outcomes and disjoint events. The probability that one of the disjoint events A or B occurs is the sum of the separate probabilities:

$$P(A \text{ or } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

GUIDED PRACTICE 3.10

(a) Verify the probability of event A, $P(A)$, is $1/3$ using the Addition Rule. (b) Do the same for event B.[4]

GUIDED PRACTICE 3.11

(a) Using Figure 3.2 as a reference, what outcomes are represented by event D? (b) Are events B and D disjoint? (c) Are events A and D disjoint?[5]

GUIDED PRACTICE 3.12

In Guided Practice 3.11, you confirmed B and D from Figure 3.2 are disjoint. Compute the probability that either event B or event D occurs.[6]

3.1.4   Probabilities when events are not disjoint

Let's consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Figure 3.3. If you are unfamiliar with the cards in a regular deck, please see the footnote.[7]

2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2◇	3◇	4◇	5◇	6◇	7◇	8◇	9◇	10◇	J◇	Q◇	K◇	A◇
2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡	A♡
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

Figure 3.3: Representations of the 52 unique cards in a deck.

GUIDED PRACTICE 3.13

(a) What is the probability that a randomly selected card is a diamond? (b) What is the probability that a randomly selected card is a face card?[8]

Venn diagrams are useful when outcomes can be categorized as "in" or "out" for two or three variables, attributes, or random processes. The Venn diagram in Figure 3.4 uses a circle to represent diamonds and another to represent face cards. If a card is both a diamond and a face card, it falls into the intersection of the circles. If it is a diamond but not a face card, it will be in part of the left circle that is not in the right circle (and so on). The total number of cards that are diamonds is given by the total number of cards in the diamonds circle: $10 + 3 = 13$. The probabilities are also shown (e.g. $10/52 = 0.1923$).

[4](a) $P(A) = P(1 \text{ or } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Similarly, $P(B) = 1/3$.
[5](a) Outcomes 2 and 3. (b) Yes, events B and D are disjoint because they share no outcomes. (c) The events A and D share an outcome in common, 2, and so are not disjoint.
[6]Since B and D are disjoint events, use the Addition Rule: $P(B \text{ or } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.
[7]The 52 cards are split into four **suits**: ♣ (club), ◇ (diamond), ♡ (heart), ♠ (spade). Each suit has its 13 cards labeled: 2, 3, ..., 10, J (jack), Q (queen), K (king), and A (ace). Thus, each card is a unique combination of a suit and a label, e.g. 4♡ and J♣. The 12 cards represented by the jacks, queens, and kings are called **face cards**. The cards that are ◇ or ♡ are typically colored red while the other two suits are typically colored black.
[8](a) There are 52 cards and 13 diamonds. If the cards are thoroughly shuffled, each card has an equal chance of being drawn, so the probability that a randomly selected card is a diamond is $P(◇) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.
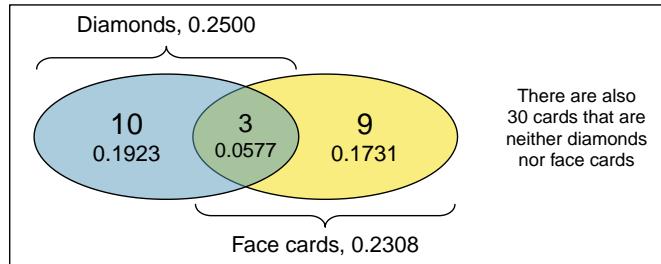
Figure 3.4: A Venn diagram for diamonds and face cards.

GUIDED PRACTICE 3.14

Using the Venn diagram, verify $P(\text{face card}) = 12/52 = 3/13$.[9]

Let A represent the event that a randomly selected card is a diamond and B represent the event that it is a face card. How do we compute $P(A \text{ or } B)$? Events A and B are not disjoint – the cards $J\diamondsuit$, $Q\diamondsuit$, and $K\diamondsuit$ fall into both categories – so we cannot use the Addition Rule for disjoint events. Instead we use the Venn diagram. We start by adding the probabilities of the two events:

$$P(A) + P(B) = P(\diamondsuit) + P(\text{face card}) = 13/52 + 12/52$$

However, the three cards that are in both events were counted twice, once in each probability. We must correct this double counting:

$$
\begin{aligned}
P(A \text{ or } B) &= P(\diamondsuit) + P(\text{face card}) \\
&= P(\diamondsuit) + P(\text{face card}) - P(\diamondsuit \text{ and face card}) \qquad (3.15) \\
&= 13/52 + 12/52 - 3/52 \\
&= 22/52 = 11/26
\end{aligned}
$$

Equation (3.15) is an example of the **General Addition Rule**.

GENERAL ADDITION RULE

If A and B are any two events, disjoint or not, then the probability that A or B will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \qquad (3.16)$$

where $P(A \text{ and } B)$ is the probability that both events occur.

SYMBOLIC NOTATION FOR "AND" AND "OR"

The symbol $\cap$ means intersection and is equivalent to "and".
The symbol $\cup$ means union and is equivalent to "or".
It is common to see the General Addition Rule written as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (3.17)$$

"OR" IS INCLUSIVE

When we write, "or" in statistics, we mean "and/or" unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur. This is equivalent to at least one of A or B occurring.

---

[9]The Venn diagram shows face cards split up into "face card but not $\diamondsuit$" and "face card and $\diamondsuit$". Since these correspond to disjoint events, $P(\text{face card})$ is found by adding the two corresponding probabilities: $\frac{3}{52} + \frac{9}{52} = \frac{12}{52} = \frac{3}{13}$.

GUIDED PRACTICE 3.18

(a) If A and B are disjoint, describe why this implies $P(A \text{ and } B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the simpler Addition Rule for disjoint events if A and B are disjoint.[10]

GUIDED PRACTICE 3.19

In the `email` data set with 3,921 emails, 367 were spam, 2,827 contained some small numbers but no big numbers, and 168 had both characteristics. Create a Venn diagram for this setup.[11]

GUIDED PRACTICE 3.20

(a) Use your Venn diagram from Guided Practice 3.19 to determine the probability a randomly drawn email from the `email` data set is spam and had small numbers (but not big numbers). (b) What is the probability that the email had either of these attributes?[12]

3.1.5 Complement of an event

Rolling a die produces a value in the set $\{1, 2, 3, 4, 5, 6\}$. This set of all possible outcomes is called the **sample space** (S) for rolling a die. We often use the sample space to examine the scenario where an event does not occur.

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D. Figure 3.5 shows the relationship between D, D^c, and the sample space S.
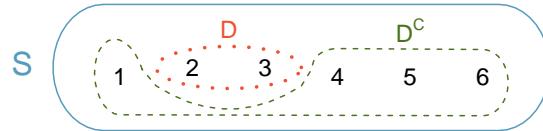
Figure 3.5: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

GUIDED PRACTICE 3.21

(a) Compute $P(D^c) = P(\text{rolling a 1, 4, 5, or 6})$. (b) What is $P(D) + P(D^c)$?[13]

---
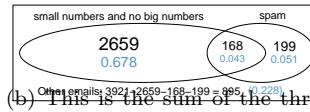
[10](a) If A and B are disjoint, A and B can never occur simultaneously. (b) If A and B are disjoint, then the last term of Equation (3.16) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

[11]Both the counts and corresponding probabilities (e.g. $2659/3921 = 0.678$) are shown. Notice that the number of emails represented in the left circle corresponds to $2659 + 168 = 2827$, and the number represented in the right circle is $168 + 199 = 367$.

[12](a) The solution is represented by the intersection of the two circles: 0.043. (b) This is the sum of the three disjoint probabilities shown in the circles: $0.678 + 0.043 + 0.051 = 0.772$.

[13](a) The outcomes are disjoint and each has probability 1/6, so the total probability is $4/6 = 2/3$. (b) We can also see that $P(D) = \frac{1}{6} + \frac{1}{6} = 1/3$. Since D and D^c are disjoint, $P(D) + P(D^c) = 1$.

GUIDED PRACTICE 3.22

Ⓖ

Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 3.2 on page 127. (a) Write out what A^c and B^c represent. (b) Compute $P(A^c)$ and $P(B^c)$. (c) Compute $P(A) + P(A^c)$ and $P(B) + P(B^c)$.[14]

An event A together with its complement A^c comprise the entire sample space. Because of this we can say that $P(A) + P(A^c) = 1$.

COMPLEMENT

The complement of event A is denoted A^c, and A^c represents all outcomes not in A. A and A^c are mathematically related:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c) \tag{3.23}$$

In simple examples, computing A or A^c is feasible in a few steps. However, using the complement can save a lot of time as problems grow in complexity.

GUIDED PRACTICE 3.24

Ⓖ

A die is rolled 10 times. (a) What is the complement of getting at least one 6 in 10 rolls of the die? (b) What is the complement of getting at most three 6's in 10 rolls of the die?[15]

3.1.6 Independence

Just as variables and observations can be independent, random processes can be independent, too. Two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

Example 3.5 provides a basic example of two independent processes: rolling two dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die. We first encountered this same question in Example 3.5 (page 125), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a 1, and $1/6^{th}$ of *those* times the white die will also be 1. This is illustrated in Figure 3.6. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6) \times (1/6) = 1/36$. This can be generalized to many independent processes.

All rolls

1/6th of the first rolls are a 1.

1/6th of those times where the first roll is a 1 the second roll is also a 1.
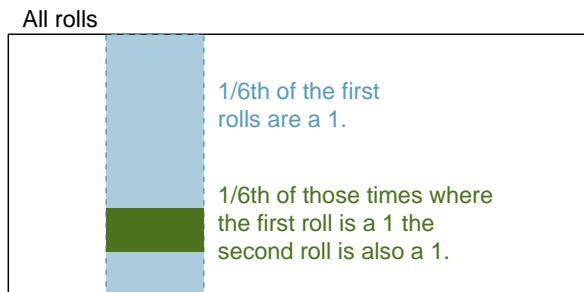
Figure 3.6: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

---

[14]Brief solutions: (a) $A^c = \{3, 4, 5, 6\}$ and $B^c = \{1, 2, 3, 5\}$. (b) Noting that each outcome is disjoint, add the individual outcome probabilities to get $P(A^c) = 2/3$ and $P(B^c) = 2/3$. (c) A and A^c are disjoint, and the same is true of B and B^c. Therefore, $P(A) + P(A^c) = 1$ and $P(B) + P(B^c) = 1$.

[15](a) The complement of getting at least one 6 in ten rolls of a die is getting zero 6's in the 10 rolls. (b) The complement of getting at most three 6's in 10 rolls is getting four, five, ..., nine, or ten 6's in 10 rolls.

EXAMPLE 3.25

What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all 1s?

The same logic applies from Example 3.5. If $1/36^{th}$ of the time the white and red dice are both 1, then $1/6^{th}$ of *those* times the blue die will also be 1, so multiply:

$$P(white = 1 \text{ and } red = 1 \text{ and } blue = 1) = P(white = 1) \times P(red = 1) \times P(blue = 1)$$
$$= (1/6) \times (1/6) \times (1/6) = 1/216$$

Examples 3.5 and 3.25 illustrate what is called the Multiplication Rule for independent processes.

MULTIPLICATION RULE FOR INDEPENDENT PROCESSES

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B) \qquad\qquad (3.26)$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

GUIDED PRACTICE 3.27

About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?[16]

GUIDED PRACTICE 3.28

Suppose 5 people are selected at random.[17]

(a) What is the probability that all are right-handed?

(b) What is the probability that all are left-handed?

(c) What is the probability that not all of the people are right-handed?

---

[16](a) The probability the first person is left-handed is 0.09, which is the same for the second person. We apply the Multiplication Rule for independent processes to determine the probability that both will be left-handed: $0.09 \times 0.09 = 0.0081$.

(b) It is reasonable to assume the proportion of people who are ambidextrous (both right- and left-handed) is nearly 0, which results in $P(\text{right-handed}) = 1 - 0.09 = 0.91$. Using the same reasoning as in part (a), the probability that both will be right-handed is $0.91 \times 0.91 = 0.8281$.

[17](a) The abbreviations RH and LH are used for right-handed and left-handed, respectively. Since each are independent, we apply the Multiplication Rule for independent processes:

$$P(\text{all five are RH}) = P(\text{first} = \text{RH}, \text{second} = \text{RH}, ..., \text{fifth} = \text{RH})$$
$$= P(\text{first} = \text{RH}) \times P(\text{second} = \text{RH}) \times \cdots \times P(\text{fifth} = \text{RH})$$
$$= 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624$$

(b) Using the same reasoning as in (a), $0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$

(c) Use the complement, $P(\text{all five are RH})$, to answer this question:

$$P(\text{not all RH}) = 1 - P(\text{all RH}) = 1 - 0.624 = 0.376$$

Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[18] using the Multiplication Rule:

$$P(\text{right-handed and female}) = P(\text{right-handed}) \times P(\text{female})$$
$$= 0.91 \times 0.50 = 0.455$$

GUIDED PRACTICE 3.29

Three people are selected at random.[19]

(a) What is the probability that the first person is male and right-handed?

(b) What is the probability that the first two people are male and right-handed?.

(c) What is the probability that the third person is female and left-handed?

(d) What is the probability that the first two people are male and right-handed and the third person is female and left-handed?

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation (3.26).

EXAMPLE 3.30

If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is $1/4$ and the probability that it is an ace is $1/13$. The probability the card is the ace of hearts is $1/52$. We check whether Equation 3.26 is satisfied:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

Because the equation holds, the event that the card is a heart and the event that the card is an ace are independent events.

---

[18]The actual proportion of the U.S. population that is `female` is about 50%, and so we use 0.5 for the probability of sampling a woman. However, this probability does differ in other countries.

[19]Brief answers are provided. (a) This can be written in probability notation as P(a randomly selected person is male and right-handed) = 0.455. (b) 0.207. (c) 0.045. (d) 0.0093.

Section summary

- When an outcome depends upon a chance process, we can define the **probability** of the outcome as the proportion of times it would occur if we repeated the process an infinite number of times. Also, even when an outcome is not truly random, modeling it with probability can be useful.

- The **Law of Large Numbers** states that the **relative frequency**, or proportion of times an outcome occurs after n repetitions, stabilizes around the true probability as n gets large.

- The probability of an event is always between 0 and 1, inclusive.

- The probability of an event and the probability of its **complement** add up to 1. Sometime we use $P(A) = 1 - P(\text{not } A)$ when $P(\text{not } A)$ is easier to calculate than $P(A)$.

- A and B are **disjoint**, i.e. **mutually exclusive**, if they cannot happen together. In this case, the events do not overlap and $P(A \text{ and } B) = 0$.

- In the *special case* where A and B are **disjoint** events: $P(A \text{ or } B) = P(A) + P(B)$.

- When A and B are not disjoint, adding $P(A)$ and $P(B)$ will overestimate $P(A \text{ or } B)$ because the overlap of A and B will be added twice. Therefore, when A and B are not disjoint, use the **General Addition Rule**:
  $P(A \text{ or B}) = P(A) + P(B) - P(A \text{ and } B)$.[20]

- To find the probability that *at least one* of several events occurs, use a special case of the rule of **complements**: $P(\text{at least one}) = 1 - P(\text{none})$.

- When only considering two events, the probability that one *or* the other happens is equal to the probability that *at least one* of the two events happens. When dealing with more than two events, the General Addition Rule becomes very complicated. Instead, to find the probability that A or B or C occurs, find the probability that none of them occur and subtract that value from 1.

- Two events are **independent** when the occurrence of one does not change the likelihood of the other.

- In the *special case* where A and B are **independent**: $P(A \text{ and } B) = P(A) \times P(B)$.

---

[20]Often written: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exercises

3.1   True or false. Determine if the statements below are true or false, and explain your reasoning.

(a) If a fair coin is tossed many times and the last eight tosses are all heads, then the chance that the next toss will be heads is somewhat less than 50%.

(b) Drawing a face card (jack, queen, or king) and drawing a red card from a full deck of playing cards are mutually exclusive events.

(c) Drawing a face card and drawing an ace from a full deck of playing cards are mutually exclusive events.

3.2   Roulette wheel. The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball.

(a) You watch a roulette wheel spin 3 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(b) You watch a roulette wheel spin 300 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(c) Are you equally confident of your answers to parts (a) and (b)? Why or why not?

Photo by Håkan Dahlström
(http://flic.kr/p/93fEzp)
CC BY 2.0 license

3.3   Four games, one winner. Below are four versions of the same game. Your archnemesis gets to pick the version of the game, and then you get to choose how many times to flip a coin: 10 times or 100 times. Identify how many coin flips you should choose for each version of the game. It costs $1 to play each game. Explain your reasoning.

(a) If the proportion of heads is larger than 0.60, you win $1.

(b) If the proportion of heads is larger than 0.40, you win $1.

(c) If the proportion of heads is between 0.40 and 0.60, you win $1.

(d) If the proportion of heads is smaller than 0.30, you win $1.

3.4   Backgammon. Backgammon is a board game for two players in which the playing pieces are moved according to the roll of two dice. Players win by removing all of their pieces from the board, so it is usually good to roll high numbers. You are playing backgammon with a friend and you roll two 6s in your first roll and two 6s in your second roll. Your friend rolls two 3s in his first roll and again in his second row. Your friend claims that you are cheating, because rolling double 6s twice in a row is very unlikely. Using probability, show that your rolls were just as likely as his.

3.5   Coin flips. If you flip a fair coin 10 times, what is the probability of

(a) getting all tails?

(b) getting all heads?

(c) getting at least one tails?

3.6   Dice rolls. If you roll a pair of fair dice, what is the probability of

(a) getting a sum of 1?

(b) getting a sum of 5?

(c) getting a sum of 12?

3.7   Swing voters. 📹 A Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.[21]

(a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of voters are Independent but not swing voters?

(d) What percent of voters are Independent or swing voters?

(e) What percent of voters are neither Independent nor swing voters?

(f) Is the event that someone is a swing voter independent of the event that someone is a political Independent?

3.8   Poverty and language. The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.[22]

(a) Are living below the poverty line and speaking a foreign language at home disjoint?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of Americans live below the poverty line and only speak English at home?

(d) What percent of Americans live below the poverty line or speak a foreign language at home?

(e) What percent of Americans live above the poverty line and only speak English at home?

(f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

3.9   Disjoint vs. independent. In parts (a) and (b), identify whether the events are disjoint, independent, or neither (events cannot be both disjoint and independent).

(a) You and a randomly selected student from your class both earn A's in this course.

(b) You and your class study partner both earn A's in this course.

(c) If two events can occur at the same time, must they be dependent?

3.10   Guessing on an exam. In a multiple choice exam, there are 5 questions and 4 choices for each question (a, b, c, d). Nancy has not studied for the exam at all and decides to randomly guess the answers. What is the probability that:

(a) the first question she gets right is the 5th question?

(b) she gets all of the questions right?

(c) she gets at least one question right?

---

[21]Pew Research Center, With Voters Focused on Economy, Obama Lead Narrows, data collected between April 4-15, 2012.

[22]U.S. Census Bureau, 2010 American Community Survey 1-Year Estimates, Characteristics of People by Language Spoken at Home.

3.11 Educational attainment of couples. The table below shows the distribution of education level attained by US residents by gender based on data collected in the 2010 American Community Survey.[23]

| | | Gender | |
		Male	Female
	Less than 9th grade	0.07	0.13
	9th to 12th grade, no diploma	0.10	0.09
Highest	HS graduate (or equivalent)	0.30	0.20
education	Some college, no degree	0.22	0.24
attained	Associate's degree	0.06	0.08
	Bachelor's degree	0.16	0.17
	Graduate or professional degree	0.09	0.09
	Total	1.00	1.00

(a) What is the probability that a randomly chosen man has at least a Bachelor's degree?

(b) What is the probability that a randomly chosen woman has at least a Bachelor's degree?

(c) What is the probability that a man and a woman getting married both have at least a Bachelor's degree? Note any assumptions you must make to answer this question.

(d) If you made an assumption in part (c), do you think it was reasonable? If you didn't make an assumption, double check your earlier answer and then return to this part.

3.12 School absences. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.[24]

(a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

(b) What is the probability that a student chosen at random misses no more than one day?

(c) What is the probability that a student chosen at random misses at least one day?

(d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

(e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

(f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

---

[23]U.S. Census Bureau, 2010 American Community Survey 1-Year Estimates, Educational Attainment.

[24]S.S. Mizan et al. "Absence, Extended Absence, and Repeat Tardiness Related to Asthma Status among Elementary School Children". In: *Journal of Asthma* 48.3 (2011), pp. 228–234.

3.2   Conditional probability

In this section we will use conditional probabilities to answer the following questions:

- What is the likelihood that a machine learning algorithm will misclassify a photo as being about fashion if it is not actually about fashion?

- How much more likely are children to attend college whose parents attended college than children whose parents did not attend college?

- Given that a person receives a positive test result for a disease, what is the probability that the person actually has the disease?

Learning objectives

1. Understand conditional probability and how to calculate it.
2. Calculate joint and conditional probabilities based on a two-way table.
3. Use the General Multiplication Rule to find the probability of joint events.
4. Determine whether two events are independent and whether they are mutually exclusive, based on the definitions of those terms.
5. Draw a tree diagram with at least two branches to organize possible outcomes and their probabilities. Understand that the second branch represents conditional probabilities.
6. Use the tree diagram or Bayes' Theorem to solve "inverted" conditional probabilities.

3.2.1 Exploring probabilities with a contingency table

The `photo_classify` data set represents a sample of 1822 photos from a photo sharing website. Data scientists have been working to improve a classifier for whether the photo is about fashion or not, and these 659 photos represent a test for their classifier. Each photo gets two classifications: the first is called `mach_learn` and gives a classification from a machine learning (ML) system of either `pred_fashion` or `pred_not`. Each of these 1822 photos have also been classified carefully by a team of people, which we take to be the source of truth; this variable is called `truth` and takes values `fashion` and `not`. Figure 3.7 summarizes the results.

		truth		
		fashion	not	Total
`mach_learn`	`pred_fashion`	197	22	219
	`pred_not`	112	1491	1603
	Total	309	1513	1822

Figure 3.7: Contingency table summarizing the `photo_classify` data set.
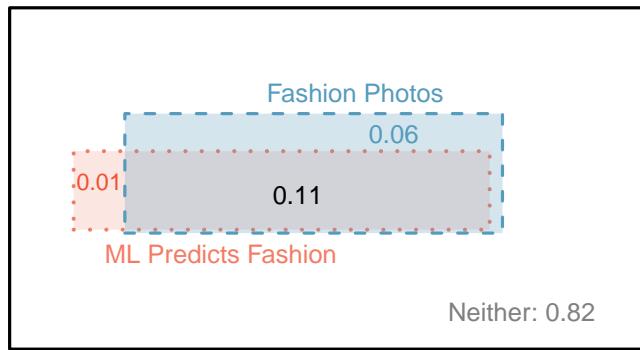
Figure 3.8: A Venn diagram using boxes for the `photo_classify` data set.

EXAMPLE 3.31

If a photo is actually about fashion, what is the chance the ML classifier correctly identified the photo as being about fashion?

We can estimate this probability using the data. Of the 309 fashion photos, the ML algorithm correctly classified 197 of the photos:

$$P(\texttt{mach\_learn is pred\_fashion given truth is fashion}) = \frac{197}{309} = 0.638$$

EXAMPLE 3.32

We sample a photo from the data set and learn the ML algorithm predicted this photo was not about fashion. What is the probability that it was incorrect and the photo is about fashion?

If the ML classifier suggests a photo is not about fashion, then it comes from the second row in the data set. Of these 1603 photos, 112 were actually about fashion:

$$P(\texttt{truth is fashion given mach\_learn is pred\_not}) = \frac{112}{1603} = 0.070$$

3.2.2   Marginal and joint probabilities

Figure 3.7 includes row and column totals for each variable separately in the `photo_classify` data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without regard to any other variables. For instance, a probability based solely on the `mach_learn` variable is a marginal probability:

$$P(\texttt{mach\_learn} \text{ is } \texttt{pred\_fashion}) = \frac{219}{1822} = 0.12$$

A probability of outcomes for two or more variables or processes is called a **joint probability**:

$$P(\texttt{mach\_learn} \text{ is } \texttt{pred\_fashion} \text{ and } \texttt{truth} \text{ is } \texttt{fashion}) = \frac{197}{1822} = 0.11$$

It is common to substitute a comma for "and" in a joint probability, although using either the word "and" or a comma is acceptable:

$$P(\texttt{mach\_learn} \text{ is } \texttt{pred\_fashion}, \texttt{truth} \text{ is } \texttt{fashion})$$

means the same thing as

$$P(\texttt{mach\_learn} \text{ is } \texttt{pred\_fashion} \text{ and } \texttt{truth} \text{ is } \texttt{fashion})$$

MARGINAL AND JOINT PROBABILITIES

If a probability is based on a single variable, it is a *marginal probability*. The probability of outcomes for two or more variables or processes is called a *joint probability*.

We use **table proportions** to summarize joint probabilities for the `photo_classify` sample. These proportions are computed by dividing each count in Figure 3.7 by the table's total, 1822, to obtain the proportions in Figure 3.9. The joint probability distribution of the `mach_learn` and `truth` variables is shown in Figure 3.10.

	truth: fashion	truth: not	Total
mach_learn: pred_fashion	0.1081	0.0121	0.1202
mach_learn: pred_not	0.0615	0.8183	0.8798
Total	0.1696	0.8304	1.00

Figure 3.9: Probability table summarizing the `photo_classify` data set.

Joint outcome	Probability
mach_learn is pred_fashion and truth is fashion	0.1081
mach_learn is pred_fashion and truth is not	0.0121
mach_learn is pred_not and truth is fashion	0.0615
mach_learn is pred_not and truth is not	0.8183
Total	1.0000

Figure 3.10: Joint probability distribution for the `photo_classify` data set.

GUIDED PRACTICE 3.33

Ⓖ

Verify Figure 3.10 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[25]

---

[25]Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.1081 + 0.0121 + 0.0615 + 0.8183 = 1.00$.

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability that a randomly selected photo from the data set is about fashion is found by summing the outcomes in which truth takes value fashion:

$$P(\underline{\text{truth is fashion}}) = P(\text{mach\_learn is pred\_fashion and } \underline{\text{truth is fashion}})$$
$$+ P(\text{mach\_learn is pred\_not and } \underline{\text{truth is fashion}})$$
$$= 0.1081 + 0.0615$$
$$= 0.1696$$

3.2.3 Defining conditional probability

The ML classifier predicts whether a photo is about fashion, even if it is not perfect. We would like to better understand how to use information from a variable like mach_learn to improve our probability estimation of a second variable, which in this example is truth.

The probability that a random photo from the data set is about fashion is about 0.17. If we knew the machine learning classifier predicted the photo was about fashion, could we get a better estimate of the probability the photo is actually about fashion? Absolutely. To do so, we limit our view to only those 219 cases where the ML classifier predicted that the photo was about fashion and look at the fraction where the photo was actually about fashion:

$$P(\text{truth is fashion given mach\_learn is pred\_fashion}) = \frac{197}{219} = 0.900$$

We call this a **conditional probability** because we computed the probability under a condition: the ML classifier prediction said the photo was about fashion.

There are two parts to a conditional probability, the **outcome of interest** and the **condition**. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event. We generally separate the text inside our probability notation into the outcome of interest and the condition with a vertical bar:

$$P(\text{truth is fashion given mach\_learn is pred\_fashion})$$
$$= P(\text{truth is fashion} \mid \text{mach\_learn is pred\_fashion}) = \frac{197}{219} = 0.900$$

The vertical bar "|" is read as *given*.

In the last equation, we computed the probability a photo was about fashion based on the condition that the ML algorithm predicted it was about fashion as a fraction:

$$P(\text{truth is fashion} \mid \text{mach\_learn is pred\_fashion})$$
$$= \frac{\text{\# cases where truth is fashion and mach\_learn is pred\_fashion}}{\text{\# cases where mach\_learn is pred\_fashion}}$$
$$= \frac{197}{219} = 0.900$$

We considered only those cases that met the condition, mach_learn is pred_fashion, and then we computed the ratio of those cases that satisfied our outcome of interest, photo was actually about fashion.

Frequently, marginal and joint probabilities are provided instead of count data. For example, disease rates are commonly listed in percentages rather than in a count format. We would like to be able to compute conditional probabilities even when no counts are available, and we use the last equation as a template to understand this technique.

We considered only those cases that satisfied the condition, where the ML algorithm predicted fashion. Of these cases, the conditional probability was the fraction representing the outcome of interest, that the photo was about fashion. Suppose we were provided only the information in Figure 3.9, i.e. only probability data. Then if we took a sample of 1000 photos, we would anticipate about 12.0% or $0.120 \times 1000 = 120$ would be predicted to be about fashion (mach_learn is pred_fashion). Similarly, we would expect about 10.8% or $0.108 \times 1000 = 108$ to meet both the in-

formation criteria and represent our outcome of interest. Then the conditional probability can be computed as

$$P(\text{truth is } \texttt{fashion} \mid \texttt{mach\_learn is pred\_fashion})$$

$$= \frac{\#\ (\text{truth is } \texttt{fashion} \text{ and } \texttt{mach\_learn is pred\_fashion})}{\#\ (\texttt{mach\_learn is pred\_fashion})}$$

$$= \frac{108}{120} = \frac{0.108}{0.120} = 0.90$$

Here we are examining exactly the fraction of two probabilities, 0.108 and 0.120, which we can write as

$$P(\text{truth is } \texttt{fashion} \text{ and } \texttt{mach\_learn is pred\_fashion}) \quad \text{and} \quad P(\texttt{mach\_learn is pred\_fashion}).$$

The fraction of these probabilities is an example of the general formula for conditional probability.

CONDITIONAL PROBABILITY

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

GUIDED PRACTICE 3.34

(a) Write out the following statement in conditional probability notation: "*The probability that the ML prediction was correct, if the photo was about fashion*". Here the condition is now based on the photo's `truth` status, not the ML algorithm.

(b) Determine the probability from part (a). Figure 3.9 on page 140 may be helpful.[26]

GUIDED PRACTICE 3.35

(a) Determine the probability that the algorithm is incorrect if it is known the photo is about fashion.

(b) Using the answers from part (a) and Guided Practice 3.34(b), compute

$$P(\texttt{mach\_learn is pred\_fashion} \mid \text{truth is } \texttt{fashion})$$
$$+\ P(\texttt{mach\_learn is pred\_not} \mid \text{truth is } \texttt{fashion})$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[27]

---

[26](a) If the photo is about fashion and the ML algorithm prediction was correct, then the ML algorithm my have a value of `pred_fashion`:

$$P(\texttt{mach\_learn is pred\_fashion} \mid \text{truth is } \texttt{fashion})$$

(b)   The   equation   for   conditional   probability   indicates   we   should   first   find $P(\texttt{mach\_learn is pred\_fashion} \text{ and } \text{truth is } \texttt{fashion}) = 0.1081$ and $P(\text{truth is } \texttt{not}) = 0.1696$.   Then the ratio represents the conditional probability: $0.1081/0.1696 = 0.6374$.

[27](a) This probability is $\frac{P(\texttt{mach\_learn is pred\_not, truth is fashion})}{P(\text{truth is fashion})} = \frac{0.0615}{0.1696} = 0.3626$. (b) The total equals 1. (c) Under the condition the photo is about fashion, the ML algorithm must have either predicted it was about fashion or predicted it was not about fashion. The complement still works for conditional probabilities, provided the probabilities are conditioned on the same information.

3.2.4 Smallpox in Boston, 1721

The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[28] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: inoculated and result. The variable inoculated takes two levels: yes or no, indicating whether the person was inoculated or not. The variable result has outcomes lived or died. These data are summarized in Tables 3.11 and 3.12.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Figure 3.11: Contingency table for the smallpox data set.

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Figure 3.12: Table proportions for the smallpox data, computed by dividing each count by the table total, 6224.

GUIDED PRACTICE 3.36

Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[29]

GUIDED PRACTICE 3.37

Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 3.36?[30]

GUIDED PRACTICE 3.38

The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone lived or died and also affect whether that person was inoculated?[31]

---

[28]Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6)*. Geneva: World Health Organization. ISBN 92-4-156110-6.

[29]$P(\texttt{result = died} \mid \texttt{not inoculated}) = \frac{P(\texttt{result = died and not inoculated})}{P(\texttt{not inoculated})} = \frac{0.1356}{0.9608} = 0.1411$.

[30]$P(\texttt{died} \mid \texttt{inoculated}) = \frac{P(\texttt{died and inoculated})}{P(\texttt{inoculated})} = \frac{0.0010}{0.0392} = 0.0255$. The death rate for individuals who were inoculated is only about 1 in 40 while the death rate is about 1 in 7 for those who were not inoculated.

[31]Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care, so income may play a role. There are other valid answers for part (c).

3.2.5 General multiplication rule

Section 3.1.6 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

GENERAL MULTIPLICATION RULE

If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

For the term $P(A|B)$, it is useful to think of A as the outcome of interest and B as the condition.

This General Multiplication Rule is simply a rearrangement of the definition for conditional probability.

EXAMPLE 3.39

Consider the `smallpox` data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

———

We will compute our answer using the General Multiplication Rule and then verify it using Figure 3.12. We want to determine

$$P(\texttt{lived} \text{ and } \texttt{not inoculated})$$

and we are given that

$$P(\texttt{lived} \mid \texttt{not inoculated}) = 0.8588$$
$$P(\texttt{not inoculated}) = 0.9608$$

Among the 96.08% of people who were not inoculated, 85.88% survived:

$$P(\texttt{lived} \text{ and } \texttt{not inoculated}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Figure 3.12 at the intersection of `no` and `lived` (with a small rounding error).

GUIDED PRACTICE 3.40

Use $P(\texttt{inoculated}) = 0.0392$ and $P(\texttt{lived} \mid \texttt{inoculated}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[32]

GUIDED PRACTICE 3.41

If 97.54% of the inoculated people lived, what proportion of inoculated people must have died?[33]

GUIDED PRACTICE 3.42

Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[34]

---

[32]The answer is 0.0382, which can be verified using Figure 3.12.

[33]There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.54% = 2.46% of the people who were inoculated died.

[34]The samples are large relative to the difference in death rates for the "inoculated" and "not inoculated" groups, so it seems there is an association between `inoculated` and `outcome`. However, as noted in the solution to Guided Practice 3.38, this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

3.2.6 Sampling without replacement

EXAMPLE 3.43

(E)

Professors sometimes select a student at random to answer a question. If each student has an equal chance of being selected and there are 15 people in your class, what is the chance that she will pick you for the next question?

If there are 15 people to ask and none are skipping class, then the probability is 1/15, or about 0.067.

EXAMPLE 3.44

If the professor asks 3 questions, what is the probability that you will not be selected? Assume that she will not pick the same person twice in a given lecture.

(E)

For the first question, she will pick someone else with probability 14/15. When she asks the second question, she only has 14 people who have not yet been asked. Thus, if you were not picked on the first question, the probability you are again not picked is 13/14. Similarly, the probability you are again not picked on the third question is 12/13, and the probability of not being picked for any of the three questions is

$$P(\text{not picked in 3 questions})$$
$$= P(\texttt{Q1} = \texttt{not\_picked}, \texttt{Q2} = \texttt{not\_picked}, \texttt{Q3} = \texttt{not\_picked}.)$$
$$= \frac{14}{15} \times \frac{13}{14} \times \frac{12}{13} = \frac{12}{15} = 0.80$$

GUIDED PRACTICE 3.45

(G)

What rule permitted us to multiply the probabilities in Example 3.44?[35]

EXAMPLE 3.46

Suppose the professor randomly picks without regard to who she already selected, i.e. students can be picked more than once. What is the probability that you will not be picked for any of the three questions?

Each pick is independent, and the probability of not being picked for any individual question is 14/15. Thus, we can use the Multiplication Rule for independent processes.

(E)

$$P(\text{not picked in 3 questions})$$
$$= P(\texttt{Q1} = \texttt{not\_picked}, \texttt{Q2} = \texttt{not\_picked}, \texttt{Q3} = \texttt{not\_picked}.)$$
$$= \frac{14}{15} \times \frac{14}{15} \times \frac{14}{15} = 0.813$$

You have a slightly higher chance of not being picked compared to when she picked a new person for each question. However, you now may be picked more than once.

---

[35] The three probabilities we computed were actually one marginal probability, $P(\texttt{Q1=not\_picked})$, and two conditional probabilities:

$$P(\texttt{Q2} = \texttt{not\_picked} \mid \texttt{Q1} = \texttt{not\_picked}) \qquad P(\texttt{Q3} = \texttt{not\_picked} \mid \texttt{Q1} = \texttt{not\_picked}, \texttt{Q2} = \texttt{not\_picked})$$

Using the General Multiplication Rule, the product of these three probabilities is the probability of not being picked in 3 questions.

GUIDED PRACTICE 3.47

Ⓖ

Under the setup of Example 3.46, what is the probability of being picked to answer all three questions?[36]

If we sample from a small population **without replacement**, we no longer have independence between our observations. In Example 3.44, the probability of not being picked for the second question was conditioned on the event that you were not picked for the first question. In Example 3.46, the professor sampled her students **with replacement**: she repeatedly sampled the entire class without regard to who she already picked.

GUIDED PRACTICE 3.48

Ⓖ

Your department is holding a raffle. They sell 30 tickets and offer seven prizes. (a) They place the tickets in a hat and draw one for each prize. The tickets are sampled without replacement, i.e. the selected tickets are not placed back in the hat. What is the probability of winning a prize if you buy one ticket? (b) What if the tickets are sampled with replacement?[37]

GUIDED PRACTICE 3.49

Ⓖ

Compare your answers in Guided Practice 3.48. How much influence does the sampling method have on your chances of winning a prize?[38]

Had we repeated Guided Practice 3.48 with 300 tickets instead of 30, we would have found something interesting: the results would be nearly identical. The probability would be 0.0233 without replacement and 0.0231 with replacement.

SAMPLING WITHOUT REPLACEMENT

When the sample size is only a small fraction of the population (under 10%), observations can be considered independent even when sampling without replacement.

---

[36] $P(\text{being picked to answer all three questions}) = \left(\frac{1}{15}\right)^3 = 0.00030$.

[37] (a) First determine the probability of not winning. The tickets are sampled without replacement, which means the probability you do not win on the first draw is 29/30, 28/29 for the second, ..., and 23/24 for the seventh. The probability you win no prize is the product of these separate probabilities: 23/30. That is, the probability of winning a prize is $1 - 23/30 = 7/30 = 0.233$. (b) When the tickets are sampled with replacement, there are seven independent draws. Again we first find the probability of not winning a prize: $(29/30)^7 = 0.789$. Thus, the probability of winning (at least) one prize when drawing with replacement is 0.211.

[38] There is about a 10% larger chance of winning a prize when using sampling without replacement. However, at most one prize may be won under this sampling procedure.

3.2.7 Independence considerations in conditional probability

If two processes are independent, then knowing the outcome of one should provide no information about the other. We can show this is mathematically true using conditional probabilities.

GUIDED PRACTICE 3.50

Let X and Y represent the outcomes of rolling two dice. (a) What is the probability that the first die, X, is 1? (b) What is the probability that both X and Y are 1? (c) Use the formula for conditional probability to compute $P(Y = 1\,|X = 1)$. (d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.[39]

We can show in Guided Practice 3.50(c) that the conditioning information has no influence by using the Multiplication Rule for independence processes:

$$
\begin{aligned}
P(Y = 1|X = 1) &= \frac{P(Y = 1 \text{ and } X = 1)}{P(X = 1)} \\
&= \frac{P(Y = 1) \times P(X = 1)}{P(X = 1)} \\
&= P(Y = 1)
\end{aligned}
$$

GUIDED PRACTICE 3.51

Ron is watching a roulette table in a casino and notices that the last five outcomes were `black`. He figures that the chances of getting `black` six times in a row is very small (about $1/64$) and puts his paycheck on red. What is wrong with his reasoning?[40]

3.2.8 Checking for independent and mutually exclusive events

If A and B are independent events, then the probability of A being true is unchanged if B is true. Mathematically, this is written as

$$
P(A|B) = P(A)
$$

The General Multiplication Rule states that $P(A \text{ and } B)$ equals $P(A|B) \times P(B)$. If A and B are independent events, we can replace $P(A|B)$ with $P(A)$ and the following multiplication rule applies:

$$
P(A \text{ and } B) = P(A) \times P(B)
$$

CHECKING WHETHER TWO EVENTS ARE INDEPENDENT

When checking whether two events A and B are independent, verify one of the following equations holds (there is no need to check both equations):

$$
P(A|B) = P(A) \qquad\qquad P(A \text{ and } B) = P(A) \times P(B)
$$

If the equation that is checked holds true (the left and right sides are equal), A and B are independent. If the equation does not hold, then A and B are dependent.

---

[39] Brief solutions: (a) 1/6. (b) 1/36. (c) $\frac{P(Y= 1 \text{ and } X= 1)}{P(X= 1)} = \frac{1/36}{1/6} = 1/6$. (d) The probability is the same as in part (c): $P(Y = 1) = 1/6$. The probability that $Y = 1$ was unchanged by knowledge about X, which makes sense as X and Y are independent.

[40] He has forgotten that the next roulette spin is independent of the previous spins. Casinos do employ this practice; they post the last several outcomes of many betting games to trick unsuspecting gamblers into believing the odds are in their favor. This is called the **gambler's fallacy**.

EXAMPLE 3.52

Are teenager college attendance and parent college degrees independent or dependent? Figure 3.13 may be helpful.

We'll use the first equation above to check for independence. If the `teen` and `parents` variables are independent, it must be true that

$$P(\texttt{teen college} \mid \texttt{parent degree}) = P(\texttt{teen college})$$

Using Figure 3.13, we check whether equality holds in this equation.

$$P(\texttt{teen college} \mid \texttt{parent degree}) \stackrel{?}{=} P(\texttt{teen college})$$
$$0.83 \neq 0.56$$

The value 0.83 came from a probability calculation using Figure 3.13: $\frac{231}{280} \approx 0.83$. Because the sides are not equal, teenager college attendance and parent degree are dependent. That is, we estimate the probability a teenager attended college to be higher if we know that one of the teen's parents has a college degree.

		parents		Total
		degree	not	
teen	college	231	214	445
	not	49	298	347
	Total	280	512	792

Figure 3.13: Contingency table summarizing the `family_college` data set.

GUIDED PRACTICE 3.53

Use the second equation in the box above to show that teenager college attendance and parent college degrees are dependent.[41]

If A and B are mutually exclusive events, then A and B cannot occur at the same time. Mathematically, this is written as

$$P(A \text{ and } B) = 0$$

The General Addition Rule states that $P(A \text{ or } B)$ equals $P(A) + P(B) - P(A \text{ and } B)$. If A and B are mutually exclusive events, we can replace $P(A \text{ and } B)$ with 0 and the following addition rule applies:

$$P(A \text{ or } B) = P(A) + P(B)$$

---

[41]We check for equality in the following equation:

$$P(\texttt{teen college, parent degree}) \stackrel{?}{=} P(\texttt{teen college}) \times P(\texttt{parent degree})$$
$$\frac{231}{792} = 0.292 \neq \frac{445}{792} \times \frac{280}{792} = 0.199$$

These terms are not equal, which confirms what we learned in Example 3.52: teenager college attendance and parent college degrees are dependent.

CHECKING WHETHER TWO EVENTS ARE MUTUALLY EXCLUSIVE (DISJOINT)

If A and B are mutually exclusive events, then they cannot occur at the same time. If asked to determine if events A and B are mutually exclusive, verify one of the following equations holds (there is no need to check both equations):

$$P(A \text{ and } B) = 0 \qquad\qquad P(A \text{ or } B) = P(A) + P(B)$$

If the equation that is checked holds true (the left and right sides are equal), A and B are mutually exclusive. If the equation does not hold, then A and B are not mutually exclusive.

EXAMPLE 3.54

Are teen college attendance and parent college degrees mutually exclusive?

———————

Looking in the table, we see that there are 231 instances where both the teenager attended college and parents have a degree, indicating the probability of both events occurring is greater than 0. Since we have found an example where both of these events happen together, these two events are not mutually exclusive. We could more formally show this by computing the probability both events occur at the same time:

$$P(\texttt{teen college}, \texttt{parent degree}) = \frac{231}{792} \neq 0$$

Since this probability is not zero, teenager college attendance and parent college degrees are not mutually exclusive.

MUTUALLY EXCLUSIVE AND INDEPENDENT ARE DIFFERENT

If two events are mutually exclusive, then if one is true, the other cannot be true. This implies the two events are in some way connected, meaning they must be dependent.
If two events are independent, then if one occurs, it is still possible for the other to occur, meaning the events are not mutually exclusive.

DEPENDENT EVENTS NEED NOT BE MUTUALLY EXCLUSIVE.

If two events are dependent, we can<u>not</u> simply conclude they are mutually exclusive. For example, the college attendance of teenagers and a college degree by one of their parents are dependent, but those events are not mutually exclusive.

3.2.9   Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

The `smallpox` data fit this description. We see the population as split by `inoculation: yes` and `no`. Following this split, survival rates were observed for each group. This structure is reflected in the tree diagram shown in Figure 3.14. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

Inoculated   Result

lived, 0.9754

yes, 0.0392   0.0392*0.9754 = 0.03824

died, 0.0246

0.0392*0.0246 = 0.00096

lived, 0.8589

no, 0.9608   0.9608*0.8589 = 0.82523

died, 0.1411

0.9608*0.1411 = 0.13557
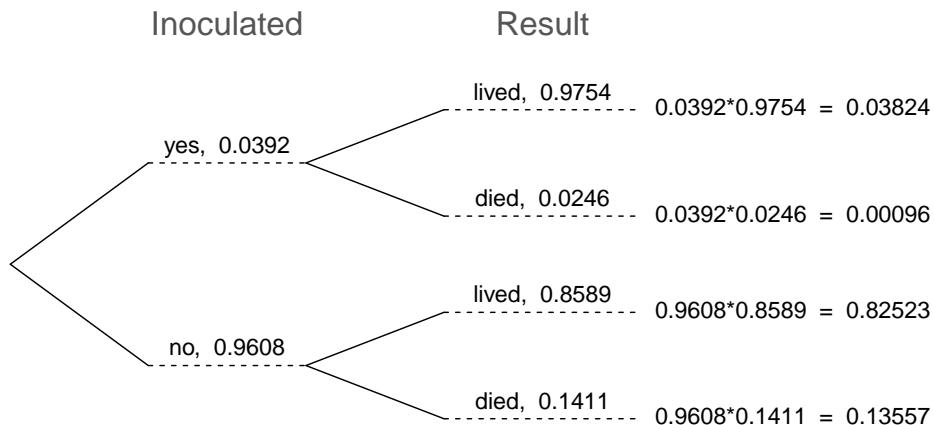
Figure 3.14: A tree diagram of the `smallpox` data set.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 3.14. This tree diagram splits the smallpox data by `inoculation` into the `yes` and `no` groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 3.14 is the probability that `lived` conditioned on the information that `inoculated`.

We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$P(\texttt{inoculated and lived}) = P(\texttt{inoculated}) \times P(\texttt{lived} \mid \texttt{inoculated})$$
$$= 0.0392 \times 0.9754$$
$$= 0.0382$$

EXAMPLE 3.55

What is the probability that a randomly selected person who was inoculated died?

This is equivalent to $P(\texttt{died} \mid \texttt{inoculated})$. This conditional probability can be found in the second branch as 0.0246.

EXAMPLE 3.56

What is the probability that a randomly selected person lived?

Ⓔ

There are two ways that a person could have lived: be inoculated *and* live OR not be inoculated *and* live. To find this probability, we sum the two disjoint probabilities:

$$P(\texttt{lived}) = 0.0392 \times 0.9745 + 0.9608 \times 0.8589 = 0.03824 + 0.82523 = 0.86347$$

GUIDED PRACTICE 3.57

Ⓖ

After an introductory statistics course, 78% of students can successfully construct tree diagrams. Of those who can construct tree diagrams, 97% passed, while only 57% of those students who could not construct tree diagrams passed. (a) Organize this information into a tree diagram. (b) What is the probability that a student who was able to construct tree diagrams did not pass? (c) What is the probability that a randomly selected student was able to successfully construct tree diagrams and passed? (d) What is the probability that a randomly selected student passed? [42]

3.2.10 Bayes' Theorem

In many instances, we are given a conditional probability of the form
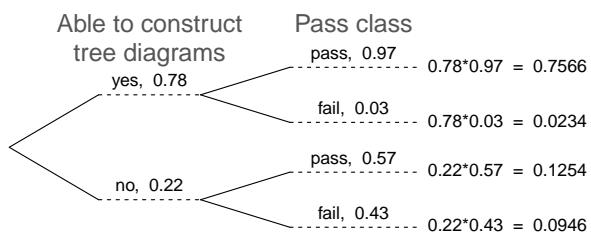
$$P(\text{statement about variable 1} \mid \text{statement about variable 2})$$

but we would really like to know the inverted conditional probability:

$$P(\text{statement about variable 2} \mid \text{statement about variable 1})$$

For example, instead of wanting to know $P(\text{lived} \mid \text{inoculated})$, we might want to know $P(\text{inoculated} \mid \text{lived})$. This is more challenging because it cannot be read directly from the tree diagram. In these instances we use **Bayes' Theorem**. Let's begin by looking at a new example.

---

[42](a) The tree diagram is shown to the right. (b) $P(\text{not pass} \mid \text{able to construct tree diagram}) = 0.03$. (c) $P(\text{able to construct tree diagrams and passed}) = P(\text{able to construct tree diagrams}) \times P(\text{passed} \mid \text{able to construct tree diagrams}) = 0.78 \times 0.97 = 0.7566$. (d) $P(\text{passed}) = 0.7566 + 0.1254 = 0.8820$.

Able to construct tree diagrams · Pass class

yes, 0.78 → pass, 0.97 → 0.78*0.97 = 0.7566
→ fail, 0.03 → 0.78*0.03 = 0.0234
no, 0.22 → pass, 0.57 → 0.22*0.57 = 0.1254
→ fail, 0.43 → 0.22*0.43 = 0.0946

EXAMPLE 3.58

In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but this test is not perfect. In about 11% of patients with breast cancer, the test gives a **false negative**: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a **false positive** in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not.[43] If we tested a random woman over 40 for breast cancer using a mammogram and the test came back positive – that is, the test suggested the patient has cancer – what is the probability that the patient actually has breast cancer?

———————

We are given sufficient information to quickly compute the probability of testing positive if a woman has breast cancer ($1.00 - 0.11 = 0.89$). However, we seek the inverted probability of cancer given a positive test result:

$$P(\text{has BC} \mid \text{mammogram}^+)$$

Here, "has BC" is an abbreviation for the patient actually having breast cancer, and "mammogram$^+$" means the mammogram screening was positive, which in this case means the test suggests the patient has breast cancer. (Watch out for the non-intuitive medical language: a *positive* test result suggests the possible presence of cancer in a mammogram screening.) We can use the conditional probability formula from the previous section: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. Our conditional probability can be found as follows:

$$P(\text{has BC} \mid \text{mammogram}^+) = \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)}$$

The probability that a mammogram is positive is as follows.

$$P(\text{mammogram}^+) = P(\text{has BC and mammogram}^+) + P(\text{no BC and mammogram}^+)$$

A tree diagram is useful for identifying each probability and is shown in Figure 3.15. Using the tree diagram, we find that

$$P(\text{has BC} \mid \text{mammogram}^+)$$
$$= \frac{P(\text{has BC and mammogram}^+)}{P(\text{has BC and mammogram}^+) + P(\text{no BC and mammogram}^+)}$$
$$= \frac{0.0035(0.89)}{0.0035(0.89) + 0.9965(0.07)}$$
$$= \frac{0.00312}{0.07288} \approx 0.0428$$

That is, even if a patient has a positive mammogram screening, there is still only a 4% chance that she has breast cancer.

Example 3.58 highlights why doctors often run more tests regardless of a first positive test result. When a medical condition is rare, a single positive test isn't generally definitive.

———————

[43] The probabilities reported here were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.
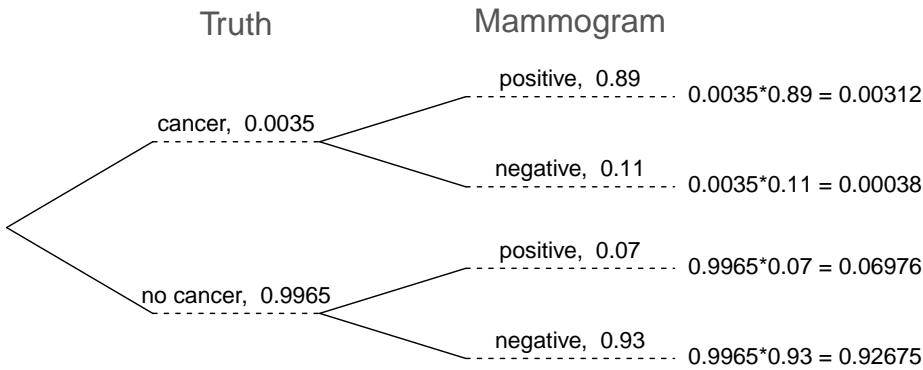
Figure 3.15: Tree diagram for Example 3.58, computing the probability a random patient who tests positive on a mammogram actually has breast cancer.

Consider again the last equation of Example 3.58. Using the tree diagram, we can see that the numerator (the top of the fraction) is equal to the following product:

$$P(\text{has BC and mammogram}^+) = P(\text{mammogram}^+|\text{ has BC})P(\text{has BC})$$

The denominator – the probability the screening was positive – is equal to the sum of probabilities for each positive screening scenario:

$$P(\underline{\text{mammogram}^+}) = P(\underline{\text{mammogram}^+}\text{ and no BC}) + P(\underline{\text{mammogram}^+}\text{ and has BC})$$

In the example, each of the probabilities on the right side was broken down into a product of a conditional probability and marginal probability using the tree diagram.

$$P(\text{mammogram}^+) = P(\text{mammogram}^+\text{ and no BC}) + P(\text{mammogram}^+\text{ and has BC})$$
$$= P(\text{mammogram}^+|\text{ no BC})P(\text{no BC})$$
$$+ P(\text{mammogram}^+|\text{ has BC})P(\text{has BC})$$

We can see an application of Bayes' Theorem by substituting the resulting probability expressions into the numerator and denominator of the original conditional probability.

$$P(\text{has BC}|\text{ mammogram}^+)$$
$$= \frac{P(\text{mammogram}^+|\text{ has BC})P(\text{has BC})}{P(\text{mammogram}^+|\text{ no BC})P(\text{no BC}) + P(\text{mammogram}^+|\text{ has BC})P(\text{has BC})}$$

BAYES' THEOREM: INVERTING PROBABILITIES

Consider the following conditional probability for variable 1 and variable 2:

$$P(\text{outcome } A_1 \text{ of variable 1}|\text{ outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable.

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The formula need not be memorized, since it can always be derived using a tree diagram:

- The numerator identifies the probability of getting both A_1 and B.

- The denominator is the overall probability of getting B. Traverse each branch of the tree diagram that ends with event B. Add up the required products.

GUIDED PRACTICE 3.59
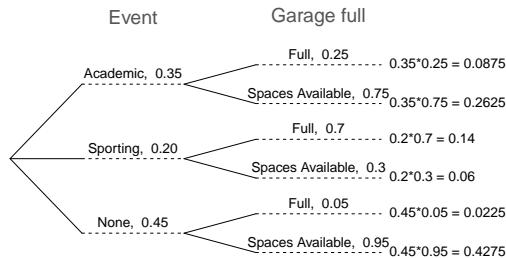
Jose visits campus every Thursday evening. However, some days the parking garage is full, often due to college events. There are academic events on 35% of evenings, sporting events on 20% of evenings, and no events on 45% of evenings. When there is an academic event, the garage fills up about 25% of the time, and it fills up 70% of evenings with sporting events. On evenings when there are no events, it only fills up about 5% of the time. If Jose comes to campus and finds the garage full, what is the probability that there is a sporting event? Use a tree diagram to solve this problem.

The tree diagram, with three primary branches, is shown to the right. We want

$$P(\text{sporting event}|\text{garage full})$$

$$= \frac{P(\text{sporting event and garage full})}{P(\text{garage full})}$$

$$= \frac{0.14}{0.0875 + 0.14 + 0.0225} = 0.56.$$

If the garage is full, there is a 56% probability that there is a sporting event.

The last several exercises offered a way to update our belief about whether there is a sporting event, academic event, or no event going on at the school based on the information that the parking lot was full. This strategy of *updating beliefs* using Bayes' Theorem is actually the foundation of an entire section of statistics called **Bayesian statistics**. While Bayesian statistics is very important and useful, we will not have time to cover it in this book.

Section summary

- A **conditional probability** can be written as $P(A|B)$ and is read, "Probability of A given B". $P(A|B)$ is the probability of A, given that B has occurred. In a conditional probability, we are given some information. In an **unconditional probability**, such as $P(A)$, we are not given any information.

- Sometimes $P(A|B)$ can be deduced. For example, when drawing without replacement from a deck of cards, $P(\text{2nd draw is an Ace} \mid \text{1st draw was an Ace}) = \frac{3}{51}$. When this is not the case, as when working with a table or a Venn diagram, one must use the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.

- In the last section, we saw that two events are **independent** when the outcome of one has no effect on the outcome of the other. When A and B are independent, $P(A|B) = P(A)$.

- When A and B are **dependent**, find the probability of A *and* B using the **General Multiplication Rule**: $P(A \text{ and } B) = P(A|B) \times P(B)$.

- In the *special case* where A and B are **independent**, $P(A \text{ and } B) = P(A) \times P(B)$.

- If A and B are **mutually exclusive**, they must be **dependent**, since the occurrence of one of them changes the probability that the other occurs to 0.

- When sampling **without replacement**, such as drawing cards from a deck, make sure to use **conditional probabilities** when solving *and* problems.

- Sometimes, the conditional probability $P(B|A)$ may be known, but we are interested in the "inverted" probability $P(A|B)$. **Bayes' Theorem** helps us solve such conditional probabilities that cannot be easily answered. However, rather than memorize Bayes' Theorem, one can generally draw a tree diagram and apply the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. The resulting answer often has the form $\frac{w \times x + y \times z}{w \times x}$, where w, x, y, z are numbers from a tree diagram.

Exercises

3.13   Joint and conditional probabilities. $P(A) = 0.3$, $P(B) = 0.7$

(a) Can you compute $P(A \text{ and } B)$ if you only know $P(A)$ and $P(B)$?

(b) Assuming that events A and B arise from independent random processes,

    i. what is $P(A \text{ and } B)$?

    ii. what is $P(A \text{ or } B)$?

    iii. what is $P(A|B)$?

(c) If we are given that $P(A \text{ and } B) = 0.1$, are the random variables giving rise to events A and B independent?

(d) If we are given that $P(A \text{ and } B) = 0.1$, what is $P(A|B)$?

3.14   PB & J. Suppose 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what's the probability that he also likes jelly?

3.15   Global warming.   A Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.[44]

		Response			
		Earth is warming	Not warming	Don't Know Refuse	Total
Party and Ideology	Conservative Republican	0.11	0.20	0.02	0.33
	Mod/Lib Republican	0.06	0.06	0.01	0.13
	Mod/Cons Democrat	0.25	0.07	0.02	0.34
	Liberal Democrat	0.18	0.01	0.01	0.20
	Total	0.60	0.34	0.06	1.00

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?

(c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

(d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

(e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

(f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

3.16 Health coverage, relative frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) and whether or not they have health insurance.

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health	No	0.0230	0.0364	0.0427	0.0192	0.0050	0.1262
Coverage	Yes	0.2099	0.3123	0.2410	0.0817	0.0289	0.8738
	Total	0.2329	0.3486	0.2838	0.1009	0.0338	1.0000

(a) Are being in excellent health and having health coverage mutually exclusive?

(b) What is the probability that a randomly chosen individual has excellent health?

(c) What is the probability that a randomly chosen individual has excellent health given that he has health coverage?

(d) What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?

(e) Do having excellent health and having health coverage appear to be independent?

3.17 Burger preferences. A 2010 SurveyUSA poll asked 500 Los Angeles residents, "What is the best hamburger place in Southern California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Or somewhere else?" The distribution of responses by gender is shown below.[45]

		Gender		
		Male	Female	Total
	Five Guys Burgers	5	6	11
	In-N-Out Burger	162	181	343
Best	Fat Burger	10	12	22
hamburger	Tommy's Hamburgers	27	27	54
place	Umami Burger	5	1	6
	Other	26	20	46
	Not Sure	13	5	18
	Total	248	252	500

(a) Are being female and liking Five Guys Burgers mutually exclusive?

(b) What is the probability that a randomly chosen male likes In-N-Out the best?

(c) What is the probability that a randomly chosen female likes In-N-Out the best?

(d) What is the probability that a man and a woman who are dating both like In-N-Out the best? Note any assumption you make and evaluate whether you think that assumption is reasonable.

(e) What is the probability that a randomly chosen person likes Umami best or that person is female?

---

[45]SurveyUSA, Results of SurveyUSA News Poll #17718, data collected on December 2, 2010.

3.18   Assortative mating. Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results. For simplicity, we only include heterosexual relationships in this exercise.[46]

| | | Partner (female) | | | |
		Blue	Brown	Green	Total
Self (male)	Blue	78	23	13	114
	Brown	19	23	12	54
	Green	11	9	16	36
	Total	108	55	41	204

(a) What is the probability that a randomly chosen male respondent or his partner has blue eyes?

(b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

(c) What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

3.19   Marbles in an urn. Imagine you have an urn containing 5 red, 3 blue, and 2 orange marbles in it.

(a) What is the probability that the first marble you draw is blue?

(b) Suppose you drew a blue marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(c) Suppose you instead drew an orange marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(d) If drawing with replacement, what is the probability of drawing two blue marbles in a row?

(e) When drawing with replacement, are the draws independent? Explain.

3.20   Socks in a drawer. In your sock drawer you have 4 blue, 5 gray, and 3 black socks. Half asleep one morning you grab 2 socks at random and put them on. Find the probability you end up wearing

(a) 2 blue socks

(b) no gray socks

(c) at least 1 black sock

(d) a green sock

(e) matching socks

3.21   Chips in a bag. Imagine you have a bag containing 5 red, 3 blue, and 2 orange chips.

(a) Suppose you draw a chip and it is blue. If drawing without replacement, what is the probability the next is also blue?

(b) Suppose you draw a chip and it is orange, and then you draw a second chip without replacement. What is the probability this second chip is blue?

(c) If drawing without replacement, what is the probability of drawing two blue chips in a row?

(d) When drawing without replacement, are the draws independent? Explain.

[46]B. Laeng et al. "Why do blue-eyed men prefer women with the same eye color?" In: *Behavioral Ecology and Sociobiology* 61.3 (2007), pp. 371–384.

3.22 Books on a bookshelf. The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

		Format		
		Hardcover	Paperback	Total
Type	Fiction	13	59	72
	Nonfiction	15	8	23
	Total	28	67	95

(a) Find the probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement.

(b) Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.

(c) Calculate the probability of the scenario in part (b), except this time complete the calculations under the scenario where the first book is placed back on the bookcase before randomly drawing the second book.

(d) The final answers to parts (b) and (c) are very similar. Explain why this is the case.

3.23 Student outfits. In a classroom with 24 students, 7 students are wearing jeans, 4 are wearing shorts, 8 are wearing skirts, and the rest are wearing leggings. If we randomly select 3 students without replacement, what is the probability that one of the selected students is wearing leggings and the other two are wearing jeans? Note that these are mutually exclusive clothing options.

3.24 The birthday problem. Suppose we pick three people at random. For each of the following questions, ignore the special case where someone might be born on February 29th, and assume that births are evenly distributed throughout the year.

(a) What is the probability that the first two people share a birthday?

(b) What is the probability that at least two people share a birthday?

3.25 Drawing box plots. After an introductory statistics course, 80% of students can successfully construct box plots. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct box plots passed.

(a) Construct a tree diagram of this scenario.

(b) Calculate the probability that a student is able to construct a box plot if it is known that he passed.

3.26 Predisposition for thrombosis. A genetic test is used to determine if people have a predisposition for *thrombosis*, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

3.27 It's never lupus. Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed tha t 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.

3.28 Exit poll. Edison Research gathered exit poll results from several sources for the Wisconsin recall election of Scott Walker. They found that 53% of the respondents voted in favor of Scott Walker. Additionally, they estimated that of those who did vote in favor for Scott Walker, 37% had a college degree, while 44% of those who voted against Scott Walker had a college degree. Suppose we randomly sampled a person who participated in the exit poll and found that he had a college degree. What is the probability that he voted in favor of Scott Walker?[47]

---

[47]New York Times, Wisconsin recall exit polls.

3.3 The binomial formula

What is the probability of exactly 50 heads in 100 coin tosses? Or the probability of randomly sampling 8 people and having exactly 1 of them be left-handed? Or of at most 3 people exceeding their insurance deductible in a random sample of 20 people? The binomial formula can help us answer these questions.

Learning objectives

1. Calculate the number of possible scenarios for obtaining x successes in n trials.

2. Determine whether a scenario is binomial or not.

3. Calculate the probability of obtaining exactly x successes in n independent trials.

4. Recognize that the binomial formula uses the special Addition Rule for mutually exclusive events.

5. Find probabilities of the form "at least" or "at most" by applying the binomial formula multiple times.

3.3.1   Introducing the binomial formula

Many health insurance plans in the United States have a deductible, where the insured individual is responsible for costs up to the deductible, and then the costs above the deductible are shared between the individual and insurance company for the remainder of the year.

Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year. Each of these people can be thought of as a **trial**. We label a person a **success** if her healthcare costs do not exceed the deductible. We label a person a **failure** if she does exceed her deductible in the year. Because 70% of the individuals will not hit their deductible, we denote the **probability of a success** as $p = 0.7$.

EXAMPLE 3.60

Suppose the insurance agency is considering a random sample of four individuals they insure. What is the chance exactly one of them will exceed the deductible and the other three will not? Let's call the four people Ariana (A), Brittany (B), Carlton (C), and Damian (D) for convenience.

Let's consider a scenario where one person exceeds the deductible:

$$P(A = \texttt{exceed},\ B = \texttt{not},\ C = \texttt{not},\ D = \texttt{not})$$
$$= P(A = \texttt{exceed})\ P(B = \texttt{not})\ P(C = \texttt{not})\ P(D = \texttt{not})$$
$$= (0.3)(0.7)(0.7)(0.7)$$
$$= (0.7)^3(0.3)^1$$
$$= 0.103$$

But there are three other scenarios: Brittany, Carlton, or Damian could have been the one to exceed the deductible. In each of these cases, the probability is again $(0.7)^3(0.3)^1$. These four scenarios exhaust all the possible ways that exactly one of these four people could have exceeded the deductible, so the total probability is $4 \times (0.7)^3(0.3)^1 = 0.412$.

GUIDED PRACTICE 3.61

Verify that the scenario where Brittany is the only one to exceed the deductible has probability $(0.7)^3(0.3)^1$. [48]

The binomial distribution describes the probability of having exactly x successes in n independent trials with probability of a success p (in Example 3.60, $n = 4$, $x = 3$, $p = 0.7$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, x, and p to obtain the probability. To do this, we reexamine each part of Example 3.60.

There were four individuals who could have been the one to exceed the deductible, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\text{\# of scenarios}] \times P(\text{single scenario})$$

The first component of this equation is the number of ways to arrange the $x = 3$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider $P(\text{single scenario})$ under the general case of x successes and $n - x$ failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^x(1 - p)^{n-x}$$

This is our general formula for $P(\text{single scenario})$.

Secondly, we introduce the **binomial coefficient**, which gives the number of ways to choose x successes in n trials, i.e. arrange x successes and $n - x$ failures:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

The quantity $\binom{n}{x}$ is read **n choose x**.[49] The exclamation point notation (e.g. $n!$) denotes a **factorial** expression.

$$0! = 1$$
$$1! = 1$$
$$2! = 2 \times 1 = 2$$
$$3! = 3 \times 2 \times 1 = 6$$
$$4! = 4 \times 3 \times 2 \times 1 = 24$$
$$\vdots$$
$$n! = n \times (n-1) \times ... \times 3 \times 2 \times 1$$

Using the formula, we can compute the number of ways to choose $x = 3$ successes in $n = 4$ trials:

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!1!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(1)} = 4$$

This result is exactly what we found by carefully thinking of each possible scenario in Example 3.60.

Substituting n choose x for the number of scenarios and $p^x(1 - p)^{n-x}$ for the single scenario probability yields the **binomial formula**.

---

[48] $P(A = \texttt{not}, \ B = \texttt{exceed}, \ C = \texttt{not}, \ D = \texttt{not}) = (0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^1$.

[49] Other notations for n choose x includes $_nC_x$, C_n^x, and $C(n, x)$.

BINOMIAL FORMULA

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

3.3.2 When and how to apply the formula

IS IT BINOMIAL? FOUR CONDITIONS TO CHECK.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

EXAMPLE 3.62

What is the probability that 3 of 8 randomly selected individuals will have exceeded the insurance deductible, i.e. that 5 of 8 will not exceed the deductible? Recall that 70% of individuals will not exceed the deductible.

We would like to apply the binomial model, so we check the conditions. The number of trials is fixed ($n = 8$) (condition 2) and each trial outcome can be classified as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are $x = 5$ successes in $n = 8$ trials (recall that a success is an individual who does *not* exceed the deductible, and the probability of a success is $p = 0.7$. So the probability that 5 of 8 will not exceed the deductible and 3 will exceed the deductible is given by

$$\binom{8}{5}(0.7)^5(1-0.7)^{8-5} = \frac{8!}{5!(5-3)!}(0.7)^5(1-0.7)^{8-5}$$

$$= \frac{8!}{5!3!}(0.7)^5(0.3)^3$$

Dealing with the factorial part:

$$\frac{8!}{5!3!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(5 \times 4 \times 3 \times 2 \times 1)(3 \times 2 \times 1)} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

Using $(0.7)^5(0.3)^3 \approx 0.00454$, the final probability is about $56 \times 0.00454 \approx 0.254$.

If you must calculate the binomial coefficient by hand, it's often useful to cancel out as many terms as possible in the top and bottom. See Section 3.3.3 for how to evaluate the binomial coefficient and the binomial formula using a calculator.

COMPUTING BINOMIAL PROBABILITIES

The first step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and x. Finally, apply the binomial formula to determine the probability and interpret the results.

EXAMPLE 3.63

Approximately 35% of a population has blood type O+. Suppose four people show up at a hospital and we want to find the probability that exactly one of them has blood type O+. Can we use the binomial formula?

———————

To check if the binomial model is appropriate, we must verify the conditions.

1. If we suppose that these 4 people comprise a random sample, then we can treat them as independent. This seems reasonable, since one person with a particular blood type showing up at a hospital seems unlikely to affect the chance that other people with that blood type would show up at the hospital.

2. We have a fixed number of trials ($n = 4$).

3. Each outcome is a success or failure (blood type O+ or not blood type O+).

4. The probability of a success is the same for each trial since the individuals are like a random sample ($p = 0.35$ if we say a "success" is someone having blood type O+).

GUIDED PRACTICE 3.64

The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3. If you have 4 friends who smoke and you want to find the probability that 1 of them will develop a severe lung condition in his or her lifetime, can you apply the binomial formula?[50]

EXAMPLE 3.65

Given that 35% of a population has blood type O+, what is the probabilty that in a random sample of 4 people:

(a) none of them have blood type O+?

(b) one will have blood type O+?

(c) no more than one will have blood type O+?

———————

Compute parts (a) and (b) using the binomial formula:

(a) $P(X = 0) = \binom{4}{0}(0.35)^0(0.65)^4 = 1 \times 1 \times 0.65^4 = 0.65^4 = 0.179$
Note that we could have answered this question without the binomial formula, using methods from the previous section.

(b) $P(X = 1) = \binom{4}{1}(0.35)^1(0.65)^3 = 0.384$.

(c) This can be computed as the sum of parts (a) and (b): $P(X = 0) + P(X = 1) = 0.179 + 0.384 = 0.563$. That is, there is about a 56.3% chance that no more than one of them will have blood type O+.

———————————————

[50]While conditions (2) and (3) are met, most likely the friends know each other, so the independence assumption (1) is probably not satisfied. For example, acquaintances may have similar smoking habits, or those friends might make a pact to quit together. Condition (4) is also not satisfied since this is not a random sample of people.

GUIDED PRACTICE 3.66

What is the probability that at least 3 of 4 people in a random sample will have blood type O+ if 35% of the population has blood type O+?[51]

EXAMPLE 3.67

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *without replacement*. Find the probability you get exactly 3 blue marbles.

_____

Because the probability of success p is not the same for each trial, we cannot use the binomial formula. However, we can use the same logic to arrive at the following answer.

$$P(X = 3) = (\# \text{ of combinations with 3 blue}) \times P(3 \text{ blue and 2 red in a } specific \text{ order})$$

$$= \binom{5}{3} \times P(\text{BBBRR})$$

$$= \binom{5}{3}\left(\frac{4}{13} \times \frac{3}{12} \times \frac{2}{11} \times \frac{9}{10} \times \frac{8}{9}\right)$$

$$= 0.1119$$

GUIDED PRACTICE 3.68

Draw 4 cards without replacement from a deck of 52 cards. What is the probability that you get at least two hearts?[52]

Lastly, we consider the binomial coefficient,, n choose x, under some special scenarios.

GUIDED PRACTICE 3.69

Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[53]

GUIDED PRACTICE 3.70

How many ways can you arrange one success and $n-1$ failures in n trials? How many ways can you arrange $n-1$ successes and one failure in n trials?[54]

_____

[51] $P(\text{at least 3 of 4 have blood type O+}) = P(X = 3) + P(X = 4) = \binom{4}{3}(0.35)^3(0.65)^1 + (0.35)^4 = 0.111 + 0.015 = 0.126$

[52] $P(\text{at least 2 hearts in 4 draws from a deck}) = 1 - [P(X = 0) + P(X = 1)] = 1 - [(\frac{39}{52})(\frac{38}{51})(\frac{37}{50})(\frac{36}{49}) + \binom{4}{1}(\frac{13}{52})(\frac{39}{51})(\frac{38}{50})(\frac{37}{49})] = 1 - [.0.3038 + 0.4388] = 0.2574.$

[53] Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

[54] One success and $n-1$ failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and $n-1$ failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

3.3.3   Calculator: binomial probabilities

TI-83/84: COMPUTING THE BINOMIAL COEFFICIENT $\binom{n}{x}$

Use `MATH`, `PRB`, `nCr` to evaluate n choose r. Here r and x are different letters for the same quantity.

1. Type the value of n.
2. Select `MATH`.
3. Right arrow to `PRB`.
4. Choose `3:nCr`.
5. Type the value of x.
6. Hit `ENTER`.

Example: `5 nCr 3` means 5 choose 3.

CASIO FX-9750GII: COMPUTING THE BINOMIAL COEFFICIENT $\binom{n}{x}$

1. Navigate to the `RUN-MAT` section (hit `MENU`, then hit `1`).
2. Enter a value for n.
3. Go to `CATALOG` (hit buttons `SHIFT` and then `7`).
4. Type `C` (hit the `ln` button), then navigate down to the bolded `C` and hit `EXE`.
5. Enter the value of x. Example of what it should look like: `7C3`.
6. Hit `EXE`.

TI-84: COMPUTING THE BINOMIAL FORMULA, $P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$

Use `2ND VARS`, `binompdf` to evaluate the probability of *exactly* x occurrences out of n independent trials of an event with probability p.

1. Select `2ND VARS` (i.e. `DISTR`)
2. Choose `A:binompdf` (use the down arrow to scroll down).
3. Let `trials` be n.
4. Let `p` be p
5. Let `x value` be x.
6. Select `Paste` and hit `ENTER`.

TI-83: Do step 1, choose `0:binompdf`, then enter n, p, and x separated by commas: `binompdf(n, p, x)`. Then hit `ENTER`.

▶ **TI-84: COMPUTING** $P(X \le x) = \binom{n}{0}p^0(1-p)^{n-0} + ... + \binom{n}{x}p^x(1-p)^{n-x}$

Use `2ND VARS`, `binomcdf` to evaluate the cumulative probability of *at most* x occurrences out of n independent trials of an event with probability p.

1. Select `2ND VARS` (i.e. `DISTR`)

2. Choose `B:binomcdf` (use the down arrow).

3. Let `trials` be n.

4. Let `p` be p

5. Let `x value` be x.

6. Select `Paste` and hit `ENTER`.

TI-83: Do steps 1-2, then enter the values for n, p, and x separated by commas as follows: `binomcdf(n, p, x)`. Then hit `ENTER`.

▶ **CASIO FX-9750GII: BINOMIAL CALCULATIONS**

1. Navigate to `STAT` (`MENU`, then hit `2`).

2. Select `DIST` (`F5`), and then `BINM` (`F5`).

3. Choose whether to calculate the binomial distribution for a specific number of successes, $P(X = k)$, or for a range $P(X \le k)$ of values (0 successes, 1 success, ..., x successes).

   - For a specific number of successes, choose `Bpd` (`F1`).
   - To consider the range 0, 1, ..., x successes, choose `Bcd`(`F1`).

4. If needed, set `Data` to `Variable` (`Var` option, which is `F2`).

5. Enter the value for `x` (x), `Numtrial` (n), and `p` (probability of a success).

6. Hit `EXE`.

GUIDED PRACTICE 3.71

(G)

Find the number of ways of arranging 3 blue marbles and 2 red marbles.[55]

GUIDED PRACTICE 3.72

(G)

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *with replacement*. Find the probability you get exactly 3 blue marbles.[56]

GUIDED PRACTICE 3.73

(G)

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *with replacement*. Find the probability you get *at most* 3 blue marbles (i.e. less than or equal to 3 blue marbles).[57]

---

[55]Here $n = 5$ and $x = 3$. Doing 5 `nCr` 3 gives the number of combinations as 10.
[56]Here, $n = 5$, $p = 4/13$, and $x = 3$, so set `trials` = 5, `p` = 4/13 and `x value` = 3. The probability is 0.1396.
[57]Similarly, set `trials` = 5, `p` = 4/13 and `x value` = 3. The cumulative probability is 0.9662.

Section summary

- $\binom{n}{x}$, the **binomial coefficient**, describes the number of combinations for arranging x successes among n trials. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where $n! = 1 \times 2 \times 3 \times ...n$, and $0!=0$.

- The **binomial formula** can be used to find the probability that something happens *exactly x times in n trials*. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

$$\binom{n}{x}p^x(1-p)^{n-x} \quad = \quad \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

- To apply the binomial formula, the events must be **independent** from trial to trial. Additionally, n, the number of trials must be fixed in advance, and p, the probability of the event occurring in a given trial, must be the same for each trial.

- To use the binomial formula, first confirm that the binomial conditions are met. Next, identify the number of trials n, the number of times the event is to be a "success" x, and the probability that a single trial is a success p. Finally, plug these three numbers into the formula to get the probability of exactly x successes in n trials.

- The $p^x(1-p)^{n-x}$ part of the binomial formula is the probability of just one combination. Since there are $\binom{n}{x}$ combinations, we add $p^x(1-p)^{n-x}$ up $\binom{n}{x}$ times. We can think of the binomial formula as: [# of combinations] $\times$ P(a single combination).

- To find a probability involving *at least* or *at most*, first determine if the scenario is binomial. If so, apply the binomial formula as many times as needed and add up the results. e.g. P(at least 3 Heads in 5 tosses of a fair coin) $=$ P(exactly 3 Heads) $+$ P(exactly 4 Heads) $+$ P(exactly 5 Heads), where each probability can be found using the binomial formula.

Exercises

3.29   Exploring combinations. A coin is tossed 5 times. How many sequences / combinations of Heads/Tails are there that have:

(a) Exactly 1 Tail?

(b) Exactly 4 Tails?

(c) Eactly 3 Tails?

(d) At least 3 Tails?

3.30   Political affiliation. Suppose that in a large population, 51% identify as Democrat. A researcher takes a random sample of 3 people.

(a) Use the binomial model to calculate the probability that two of them identify as Democrat.

(b) Write out all possible orderings of 3 people, 2 of whom identify as Democrat. Use these scenarios to calculate the same probability from part (a) but using the Addition Rule for disjoint events. Confirm that your answers from parts (a) and (b) match.

(c) If we wanted to calculate the probability that a random sample of 8 people will have 3 that identify as Democrat, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

3.31   Underage drinking, Part I. 📹 Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.7% of 18-20 year olds consumed alcoholic beverages in any given year.[58]

(a) Suppose a random sample of ten 18-20 year olds is taken. Is the use of the binomial distribution appropriate for calculating the probability that exactly six consumed alcoholic beverages? Explain.

(b) Calculate the probability that exactly 6 out of 10 randomly sampled 18- 20 year olds consumed an alcoholic drink.

(c) What is the probability that exactly four out of ten 18-20 year olds have *not* consumed an alcoholic beverage?

(d) What is the probability that at most 2 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

(e) What is the probability that at least 1 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

3.32   Chicken pox, Part I. The National Vaccine Information Center estimates that 90% of Americans have had chickenpox by the time they reach adulthood.[59]

(a) Suppose we take a random sample of 100 American adults. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood? Explain.

(b) Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.

(c) What is the probability that exactly 3 out of a new sample of 100 American adults have *not* had chickenpox in their childhood?

(d) What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?

(e) What is the probability that at most 3 out of 10 randomly sampled American adults have *not* had chickenpox?

---

[58]SAMHSA, Office of Applied Studies, National Survey on Drug Use and Health, 2007 and 2008.
[59]National Vaccine Information Center, Chickenpox, The Disease & The Vaccine Fact Sheet.

3.4 Simulations

What is the probability of getting a sum greater than 16 in three rolls of a die? Finding all possible combinations that satisfy this would be tedious, but we could conduct a physical simulation or a computer simulation to estimate this probability. With modern computing power, simulations have become an important and powerful tool for data scientists. In this section, we will look at the concepts that underlie simulations.

Learning objectives

1. Understand the purpose of a simulation and recognize the application of the long-run relative frequency interpretation of probability.

2. Understand how random digit tables work and how to assign digits to outcomes.

3. Be able to repeat a simulation a set number of trials or until a condition is true, and use the results to estimate the probability of interest.

3.4.1 Setting up and carrying out simulations

In the previous section we saw how to apply the binomial formula to find the probability of exactly x successes in n independent trials when a success has probability p. Sometimes we have a problem we want to solve but we don't know the appropriate formula, or even worse, a formula may not exist. In this case, one common approach is to estimate the probability using **simulations**.

You may already be familiar with simulations. Want to know the probability of rolling a sum of 7 with a pair of dice? Roll a pair of dice many, many, many times and see what proportion of times the sum is 7. The more times you roll the pair of dice, the better the estimate will tend to be. Of course, such experiments can be time consuming or even infeasible.

In this section, we consider simulations using **random numbers**. Random numbers (or technically, *psuedo-random numbers*) can be produced using a calculator or computer. Random digits are produced such that each digit, *0-9*, is equally likely to come up in each spot. You'll find that occasionally we may have the same number in a row – sometimes multiple times – but in the long run, each digit should appear 1/10th of the time.

| | Column | | | |
Row	1-5	6-10	11-15	16-20
1	43087	41864	51009	39689
2	63432	72132	40269	56103
3	19025	83056	62511	52598
4	85117	16706	31083	24816
5	16285	56280	01494	90240
6	94342	18473	50845	77757
7	61099	14136	39052	50235
8	37537	58839	56876	02960
9	04510	16172	90838	15210
10	27217	12151	52645	96218

Figure 3.16: Random number table. A full page of random numbers may be found in Appendix C.1 on page 519.

EXAMPLE 3.74

Mika's favorite brand of cereal is running a special where 20% of the cereal boxes contain a prize. Mika really wants that prize. If her mother buys 6 boxes of the cereal over the next few months, what is the probability Mika will get a prize?

To solve this problem using simulation, we need to be able to assign digits to outcomes. Each box should have a 20% chance of having a prize and an 80% chance of not having a prize. Therefore, a valid assignment would be:

$$0, \ 1 \rightarrow \text{prize}$$
$$2\text{-}9 \rightarrow \text{no prize}$$

Of the ten possible digits (*0, 1, 2, ..., 8, 9*), two of them, i.e. 20% of them, correspond to winning a prize, which exactly matches the odds that a cereal box contains a prize.

In Mika's simulation, one trial will consist of 6 boxes of cereal, and therefore a trial will require six digits (each digit will correspond to one box of cereal). We will repeat the simulation for 20 trials. Therefore we will need 20 sets of 6 digits. Let's begin on row 1 of the random digit table, shown in Figure 3.16. If a trial consisted of 5 digits, we could use the first 5 digits going across: *43087*. Because here a trial consists of 6 digits, it may be easier to read down the table, rather than read across. We will let trial 1 consist of the first 6 digits in column 1 (*461819*), trial 2 consist of the first 6 digits in column 2 (*339564*), etc. For this simulation, we will end up using the first 6 rows of each of the 20 columns.

In trial 1, there are two *1*'s, so we record that as a success; in this trial there were actually two prizes. In trial 2 there were no *0*'s or *1*'s, therefore we do not record this as a success. In trial 3 there were three prizes, so we record this as a success. The rest of this exercise is left as a Guided Practice problem for you to complete.

Ⓖ

GUIDED PRACTICE 3.75

Finish the simulation above and report the estimate for the probability that Mika will get a prize if her mother buys 6 boxes of cereal where each one has a 20% chance of containing a prize.[60]

Ⓖ

GUIDED PRACTICE 3.76

In the previous example, the probability that a box of cereal contains a prize is 20%. The question presented is equivalent to asking, what is the probability of getting at least one prize in six randomly selected boxes of cereal. This probability question can be solved explicitly using the method of complements. Find this probability. How does the estimate arrived at by simulation compare to this probability?[61]

We can also use simulations to estimate quantities other than probabilities. Consider the following example.

EXAMPLE 3.77

Let's say that instead of buying exactly 6 boxes of cereal, Mika's mother agrees to buy boxes of this cereal *until* she finds one with a prize. On average, how many boxes of cereal would one have to buy until one gets a prize?

Ⓔ

For this question, we can use the same digit assignment. However, our stopping rule is different. Each trial may require a different number of digits. For each trial, the stopping rule is: look at digits until we encounter a *0* or a *1*. Then, record how many digits/boxes of cereal it took. Repeat the simulation for 20 trials, and then average the numbers from each trial.

Let's begin again at row 1. We can read across or down, depending upon what is most convenient. Since there are 20 columns and we want 20 trials, we will read down the columns. Starting at column 1, we count how many digits (boxes of cereal) we encounter until we reach a *0* or *1* (which represent a prize). For trial 1 we see *461*, so we record 3. For trial 2 we see *3395641*, so we record 7. For trial 3, we see *0*, so we record 1. The rest of this exercise is left as a Guided Practice problem for you to complete.

Ⓖ

GUIDED PRACTICE 3.78

Finish the simulation above and report your estimate for the average number of boxes of cereal one would have to buy until encountering a prize, where the probability of a prize in each box is 20%.[62]

---

[60]The trials that contain at least one 0 or 1 and therefore are successes are trials: 1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, and 20. There were 17 successes among the 20 trials, so our estimate of the probability based on this simulation is $17/20 = 0.85$.

[61]The true probability is given by $1 - P(\text{no prizes in six boxes}) = 1 - 0.8^6 = 0.74$. The estimate arrived at by simulation was 11% too high. Note: We only repeated the simulation 20 times. If we had repeated it 1000 times, we would (very likely) have gotten an estimate closer to the true probability.

[62]For the 20 trials, the number of digits we see until we encounter a *0* or *1* is: 3,7,1,4,9, 4,1,2,4,5, 5,1,1,1,3, 8,5,2,2,6. Now we take the average of these 20 numbers to get $74/20 = 3.7$.

EXAMPLE 3.79

Now, consider a case where the probability of interest is not 20%, but rather 28%. Which digits should correspond to success and which to failure?

———————

This example is more complicated because with only 10 digits, there is no way to select exactly 28% of them. Therefore, each observation will have to consist of *two* digits. We can use two digits at a time and assign pairs of digits as follows:

$$00\text{-}27 \rightarrow \text{success}$$
$$28\text{-}99 \rightarrow \text{failure}$$

GUIDED PRACTICE 3.80

Assume the probability of winning a particular casino game is 45%. We want to carry out a simulation to estimate the probability that we will win at least 5 times in 10 plays. We will use 30 trials of the simulation. Assign digits to outcomes. Also, how many total digits will we require to run this simulation?[63]

GUIDED PRACTICE 3.81

Assume carnival spinner has 7 slots. We want to carry out a simulation to estimate the probability that we will win at least 10 times in 60 plays. Repeat 100 trials of the simulation. Assign digits to outcomes. Also, how many total digits will we require to run this simulation?[64]

Does anyone perform simulations like this? Sort of. Simulations are used a lot in statistics, and these often require the same principles covered in this section to properly set up those simulations. The difference is in implementation after the setup. Rather than use a random number table, a data scientist will write a program that uses a pseudo-random number generator in a computer to run the simulations very quickly – often times millions of trials each second, which provides much more accurate estimates than running a couple dozen trials by hand.

---

[63]One possible assignment is: *00-44* → win and *45-99* → lose. Each trial requires 10 pairs of digits, so we will need 30 sets of 10 pairs of digits for a total of $30 \times 10 \times 2 = 600$ digits.

[64]Note that $1/7 = 0.142857...$ This makes it tricky to assign digits to outcomes. The best approach here would be to exclude some of the digits from the simulation. We can assign *0* to success and *1-6* to failure. This corresponds to a $1/7$ chance of getting a success. If we encounter a *7*, *8*, or *9*, we will just skip over it. Because we don't know how many *7*, *8*, or *9*'s we will encounter, we do not know how many total digits we will end up using for the simulation. (If you want a challenge, try to estimate the total number of digits you would need.)

Section summary

- When a probability is difficult to determine via a formula, one can set up a **simulation** to estimate the probability.

- The **relative frequency** theory of probability and the **Law of Large Numbers** are the mathematical underpinning of simulations. A larger number of trials should tend to produce better estimates.

- The first step to setting up a simulation is to assign digits to represent outcomes. This should be done in such a way as to give the event of interest the correct probability. Then, using a random number table, calculator, or computer, generate random digits (outcomes). Repeat this a specified number of trials or until a given stopping rule. When this is finished, count up how many times the event happened and divide that by the number of trials to get the estimate of the probability.

Exercises

3.33   Smog check, Part I. Suppose 16% of cars fail pollution tests (smog checks) in California. We would like to estimate the probability that an entire fleet of seven cars would pass using a simulation. We assume each car is independent. We only want to know if the entire fleet passed, i.e. none of the cars failed. What is wrong with each of the following simulations to represent whether an entire (simulated) fleet passed?

(a) Flip a coin seven times where each toss represents a car. A head means the car passed and a tail means it failed. If all cars passed, we report PASS for the fleet. If at least one car failed, we report FAIL.

(b) Read across a random number table starting at line 5. If a number is a 0 or 1, let it represent a failed car. Otherwise the car passes. We report PASS if all cars passed and FAIL otherwise.

(c) Read across a random number table, looking at two digits for each simulated car. If a pair is in the range [00-16], then the corresponding car failed. If it is in [17-99], the car passed. We report PASS if all cars passed and FAIL otherwise.

3.34   Left-handed. Studies suggest that approximately 10% of the world population is left-handed. Use ten simulations to answer each of the following questions. For each question, describe your simulation scheme clearly.

(a) What is the probability that at least one out of eight people are left-handed?

(b) On average, how many people would you have to sample until the first person who is left-handed?

(c) On average, how many left-handed people would you expect to find among a random sample of six people?

3.35   Smog check, Part II. Consider the fleet of seven cars in Exercise 3.33. Remember that 16% of cars fail pollution tests (smog checks) in California, and that we assume each car is independent.

(a) Write out how to calculate the probability of the fleet failing, i.e. at least one of the cars in the fleet failing, via simulation.

(b) Simulate 5 fleets. Based on these simulations, estimate the probability at least one car will fail in a fleet.

(c) Compute the probability at least one car fails in a fleet of seven.

3.36   To catch a thief. Suppose that at a retail store, $1/5^{th}$ of all employees steal some amount of merchandise. The stores would like to put an end to this practice, and one idea is to use lie detector tests to catch and fire thieves. However, there is a problem: lie detectors are not 100% accurate. Suppose it is known that a lie detector has a failure rate of 25%. A thief will slip by the test 25% of the time and an honest employee will only pass 75% of the time.

(a) Describe how you would simulate whether an employee is honest or is a thief using a random number table. Write your simulation very carefully so someone else can read it and follow the directions exactly.

(b) Using a random number table, simulate 20 employees working at this store and determine if they are honest or not. Make sure to record the random digits assigned to each employee as you will refer back to these in part (c).

(c) Determine the result of the lie detector test for each simulated employee from part (b) using a new simulation scheme.

(d) How many of these employees are "honest and passed" and how many are "honest and failed"?

(e) How many of these employees are "thief and passed" and how many are "thief and failed"?

(f) Suppose the management decided to fire everyone who failed the lie detector test. What percent of fired employees were honest? What percent of not fired employees were thieves?

3.5 Random variables

The chance of landing on single number in the game of roulette is 1/38 and the pay is 35:1. The chance of landing on Red is 18/38 and the pay is 1:1. Which game has the higher expected value? The higher standard deviation of expected winnings? How do we interpret these quantities in this context? If you were to play each game 20 times, what would the *distribution* of possible outcomes look like? In this section, we define and summarize random variables such as this, and we look at some of their properties.

Learning objectives

1. Define a probability distribution and what makes a distribution a valid probability distribution.

2. Summarize a discrete probability distribution graphically using a histogram and verbally with respect to center, spread, and shape.

3. Calculate and interpret the mean (expected value) and standard deviation of a random variable.

4. Calculate the mean and standard deviation of a transformed random variable.

5. Calculate the mean of the sum or difference of random variables.

6. Calculate the standard deviation of the sum or difference of random variables when those variables are independent.

3.5.1 Introduction to expected value

EXAMPLE 3.82

Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one term to another. If there are 100 students enrolled, how many books should the bookstore expect to sell to this class?

Around 20 students will not buy either book (0 books total), about 55 will buy one book (55 books total), and approximately 25 will buy two books (totaling 50 books for these 25 students). The bookstore should expect to sell about 105 books for this class.

GUIDED PRACTICE 3.83

Would you be surprised if the bookstore sold slightly more or less than 105 books?[65]

---

[65]If they sell a little more or a little less, this should not be a surprise. Hopefully Chapter 2 helped make clear that there is natural variability in observed data. For example, if we would flip a coin 100 times, it will not usually come up heads exactly half the time, but it will probably be close.

EXAMPLE 3.84

The textbook costs \$137 and the study guide \$33. How much revenue should the bookstore expect from this class of 100 students?

About 55 students will just buy a textbook, providing revenue of

$$\$137 \times 55 = \$7,535$$

The roughly 25 students who buy both the textbook and the study guide would pay a total of

$$(\$137 + \$33) \times 25 = \$170 \times 25 = \$4,250$$

Thus, the bookstore should expect to generate about $\$7,535 + \$4,250 = \$11,785$ from these 100 students for this one class. However, there might be some *sampling variability* so the actual amount may differ by a little bit.
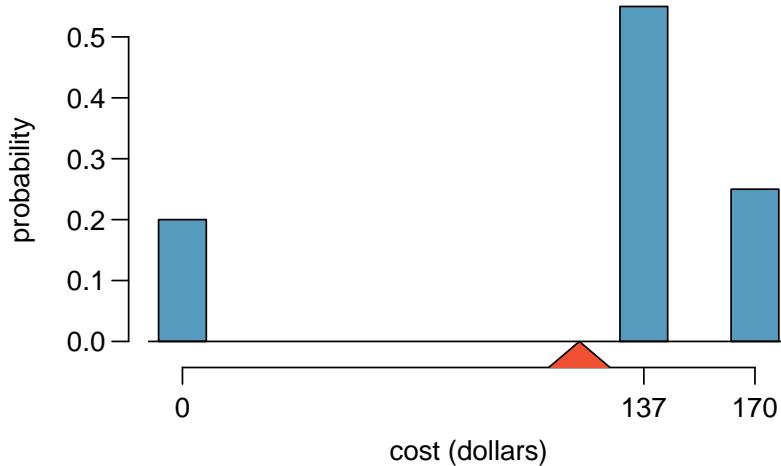
Figure 3.17: Probability distribution for the bookstore's revenue from one student. The triangle represents the average revenue per student.

EXAMPLE 3.85

What is the average revenue per student for this course?

The expected total revenue is \$11,785, and there are 100 students. Therefore the expected revenue per student is $\$11,785/100 = \$117.85$.

3.5.2 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Figure 3.18 shows the probability distribution for the sum of two dice.

RULES FOR PROBABILITY DISTRIBUTIONS

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

GUIDED PRACTICE 3.86

Ⓖ

Figure 3.19 suggests three distributions for household income in the United States. Only one is correct. Which one must it be? What is wrong with the other two?[66]

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Figure 3.18: Probability distribution for the sum of two dice.

Income range ($1000s)	0-25	25-50	50-100	100+
(a)	0.18	0.39	0.33	0.16
(b)	0.38	-0.27	0.52	0.37
(c)	0.28	0.27	0.29	0.16

Figure 3.19: Proposed distributions of US household incomes (Guided Practice 3.86).

Chapter 2 emphasized the importance of plotting data to provide quick summaries. Probability distributions can also be summarized in a histogram or bar plot. The probability distribution for the sum of two dice is shown in Figure 3.18 and its histogram is plotted in Figure 3.20. The distribution of US household incomes is shown in Figure 3.21 as a bar plot. The presence of the 100+ category makes it difficult to represent it with a regular histogram.[67]
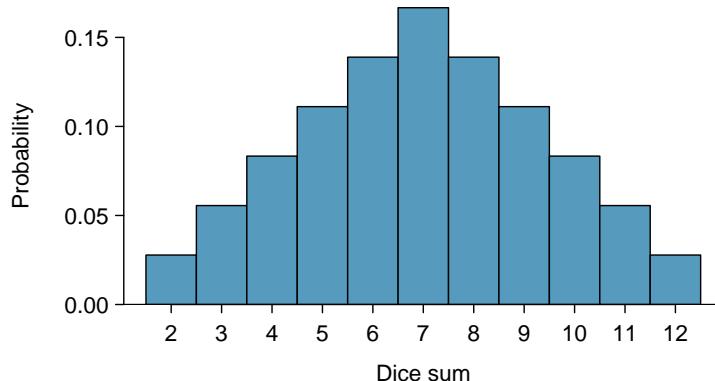
Figure 3.20: A histogram for the probability distribution of the sum of two dice.

In these bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a histogram, as in the case of the sum of two dice. Another example of plotting the bars at their respective locations is shown in Figure 3.17.

---

[66]The probabilities of (a) do not sum to 1. The second probability in (b) is negative. This leaves (c), which sure enough satisfies the requirements of a distribution. One of the three was said to be the actual distribution of US household incomes, so it must be (c).

[67]It is also possible to construct a distribution plot when income is not artificially binned into four groups. Density histograms for *continuous* distributions are considered in Section 3.6.
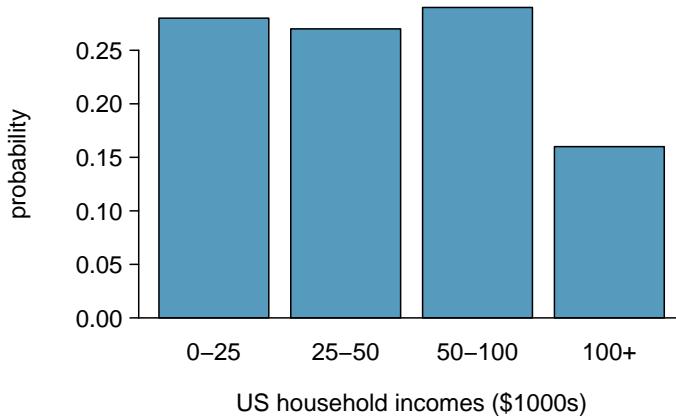
Figure 3.21: A bar graph for the probability distribution of US household income. Because it is artificially separated into four unequal bins, this graph fails to show the shape or skew of the distribution.

3.5.3 Expectation

We call a variable or process with a numerical outcome a **random variable**, and we usually represent this random variable with a capital letter such as X, Y, or Z. The amount of money a single student will spend on her statistics books is a random variable, and we represent it by X.

RANDOM VARIABLE

A random process or variable with a numerical outcome.

The possible outcomes of X are labeled with a corresponding lower case letter x and subscripts. For example, we write $x_1 = \$0$, $x_2 = \$137$, and $x_3 = \$170$, which occur with probabilities 0.20, 0.55, and 0.25. The distribution of X is summarized in Figure 3.17 and Figure 3.22.

i	1	2	3	Total
x_i	\$0	\$137	\$170	–
$P(x_i)$	0.20	0.55	0.25	1.00

Figure 3.22: The probability distribution for the random variable X, representing the bookstore's revenue from a single student. We use $P(x_i)$ to represent the probability of x_i.

We computed the average outcome of X as \$117.85 in Example 3.85. We call this average the **expected value** of X, denoted by $E(X)$. The expected value of a random variable is computed by adding each outcome weighted by its probability:

$$E(X) = 0 \cdot P(0) + 137 \cdot P(137) + 170 \cdot P(170)$$
$$= 0 \cdot 0.20 + 137 \cdot 0.55 + 170 \cdot 0.25 = 117.85$$

EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

If X takes outcomes x_1, x_2, ..., x_n with probabilities $P(x_1)$, $P(x_2)$, ..., $P(x_n)$, the mean, or expected value, of X is the sum of each outcome multiplied by its corresponding probability:

$$\mu_x = E(X) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n)$$
$$= \sum_{i=1}^{n} x_i \cdot P(x_i)$$

The expected value for a random variable represents the average outcome. For example, $E(X) = 117.85$ represents the average amount the bookstore expects to make from a single student, which we could also write as $\mu = 117.85$. While the bookstore will make more than this on some students and less than this on other students, the average of many randomly selected students will be near $117.85.

It is also possible to compute the expected value of a continuous random variable (see Section 3.6). However, it requires a little calculus and we save it for a later class.[68]

In physics, the expectation holds the same meaning as the center of gravity. The distribution can be represented by a series of weights at each outcome, and the mean represents the balancing point. This is represented in Figures 3.17 and 3.23. The idea of a center of gravity also expands to continuous probability distributions. Figure 3.24 shows a continuous probability distribution balanced atop a wedge placed at the mean.
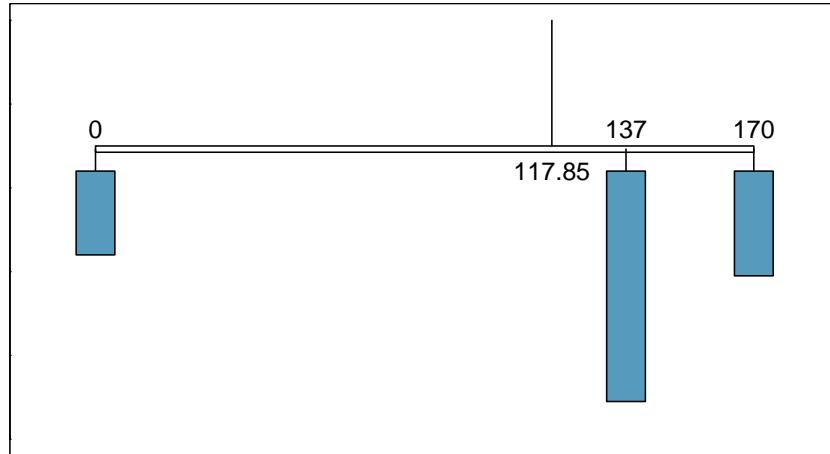
Figure 3.23: A weight system representing the probability distribution for X. The string holds the distribution at the mean to keep the system balanced.
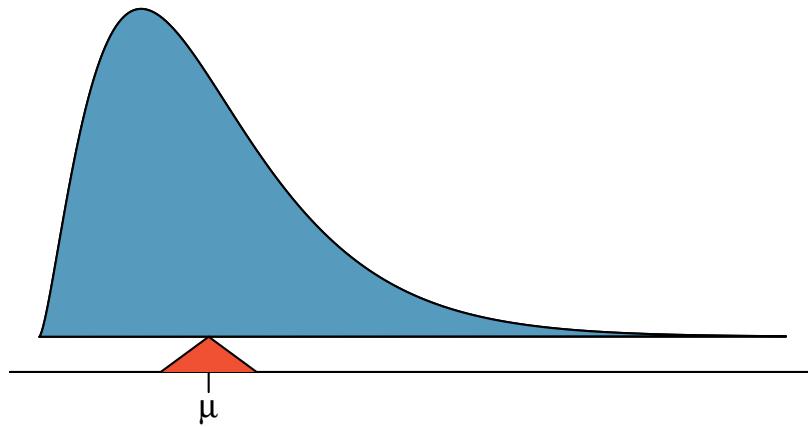
Figure 3.24: A continuous distribution can also be balanced at its mean.

---

[68]$\mu_X = \int x f(x) dx$ where $f(x)$ represents a function for the density curve.

3.5.4   Variability in random variables

Suppose you ran the university bookstore. Besides how much revenue you expect to generate, you might also want to know the volatility (variability) in your revenue.

The variance and standard deviation can be used to describe the variability of a random variable. Section 2.2.3 introduced a method for finding the variance and standard deviation for a data set. We first computed deviations from the mean $(x_i - \mu)$, squared those deviations, and took an average to get the variance. In the case of a random variable, we again compute squared deviations. However, we take their sum weighted by their corresponding probabilities, just like we did for the expectation. This weighted sum of squared deviations equals the variance, and we calculate the standard deviation by taking the square root of the variance, just as we did in Section 2.2.3.

VARIANCE AND STANDARD DEVIATION OF A DISCRETE RANDOM VARIABLE

If X takes outcomes x_1, x_2, ..., x_n with probabilities $P(x_1)$, $P(x_2)$, ..., $P(x_n)$ and expected value $\mu_X = E(X)$, then to find the standard deviation of X, we first find the variance and then take its square root.

$$Var(X) = \sigma_x^2 = (x_1 - \mu_X)^2 \cdot P(x_1) + (x_2 - \mu_X)^2 \cdot P(x_2) + \cdots + (x_n - \mu_X)^2 \cdot P(x_n)$$

$$= \sum_{i=1}^{n} (x_i - \mu_X)^2 \cdot P(x_i)$$

$$SD(X) = \sigma_X = \sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2 \cdot P(x_i)}$$

Just as it is possible to compute the mean of a continuous random variable using calculus, we can also use calculus to compute the variance.[69] However, this topic is beyond the scope of the AP exam.

EXAMPLE 3.87

Compute the expected value, variance, and standard deviation of X, the revenue of a single statistics student for the bookstore.

It is useful to construct a table that holds computations for each outcome separately, then add up the results.

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(x_i)$	0.20	0.55	0.25	
$x_i \cdot P(x_i)$	0	75.35	42.50	117.85

Thus, the expected value is $\mu_X = 117.85$, which we computed earlier. The variance can be constructed using a similar table:

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(x_i)$	0.20	0.55	0.25	
$x_i - \mu_X$	-117.85	19.15	52.15	
$(x_i - \mu_X)^2$	13888.62	366.72	2719.62	
$(x_i - \mu_X)^2 \cdot P(x_i)$	2777.7	201.7	679.9	3659.3

The variance of X is $\sigma_X^2 = 3659.3$, which means the standard deviation is $\sigma_X = \sqrt{3659.3} = \$60.49$.

---

[69]$\sigma_x^2 = \int (x - \mu_X)^2 f(x) dx$ where $f(x)$ represents a function for the density curve.

GUIDED PRACTICE 3.88

The bookstore also offers a chemistry textbook for $159 and a book supplement for $41. From past experience, they know about 25% of chemistry students just buy the textbook while 60% buy both the textbook and supplement.[70]

(a) What proportion of students don't buy either book? Assume no students buy the supplement without the textbook.

(b) Let Y represent the revenue from a single student. Write out the probability distribution of Y, i.e. a table for each outcome and its associated probability.

(c) Compute the expected revenue from a single chemistry student.

(d) Find the standard deviation to describe the variability associated with the revenue from a single student.

3.5.5   Linear transformations of a random variable

An online store is selling a limited edition t-shirt. The maximum a person is allowed to buy is 3. Let X be a random variable that represents how many of the t-shirts a t-shirt buyer orders. The probability distribution of X is given in the following table.

x_i	1	2	3
$P(x_i)$	0.6	0.3	0.1

Using the methods of the previous section we can find that the mean $\mu_X = 1.5$ and the standard deviation $\sigma_X = 0.67$. Suppose that the cost of each t-shirt is $30 and that there is flat rate $5 shipping fee. The amount of money a t-shirt buyer pays, then, is $30X + 5$, where X is the number of t-shirts ordered. To calculate the mean and standard deviation for the amount of money a t-shirt buyers pays, we could define a new variable Y as follows:

$$Y = 30X + 5$$

GUIDED PRACTICE 3.89

Verify that the distribution of Y is given by the table below.[71]

y_i	$35	$65	$95
$P(y_i)$	0.6	0.3	0.1

---

[70](a) 100% - 25% - 60% = 15% of students do not buy any books for the class. Part (b) is represented by the first two lines in the table below. The expectation for part (c) is given as the total on the line $y_i \cdot P(y_i)$. The result of part (d) is the square-root of the variance listed on in the total on the last line: $\sigma_Y = \sqrt{Var(Y)} = \sqrt{4800} = 69.28$.

i (scenario)	1 (noBook)	2 (textbook)	3 (both)	Total
y_i	0.00	159.00	200.00	
$P(y_i)$	0.15	0.25	0.60	
$y_i \cdot P(y_i)$	0.00	39.75	120.00	$E(Y) = 159.75$
$y_i - \mu_Y$	-159.75	-0.75	40.25	
$(y_i - \mu_Y)^2$	25520.06	0.56	1620.06	
$(y_i - \mu_Y)^2 \cdot P(y_i)$	3828.0	0.1	972.0	$Var(Y) \approx 4800$

[71]$30 \times 1 + 5 = 35; 30 \times 2 + 5 = 65; 30 \times 3 + 5 = 95$

Using this new table, we can compute the mean and standard deviation of the cost for t-shirt orders. However, because Y is a linear transformation of X, we can use the properties from Section 2.2.8. Recall that multiplying every X by 30 multiplies both the mean and standard deviation by 30. Adding 5 only adds 5 to the mean, not the standard deviation. Therefore,

$$
\begin{aligned}
\mu_{30X+5} &= E(30X+5) & \sigma_{30X+5} &= SD(30X+5) \\
&= 30 \times E(X) + 5 & &= 30 \times SD(X) \\
&= 30 \times 1.5 + 5 & &= 30 \times 0.67 \\
&= 45.00 & &= 20.10
\end{aligned}
$$

Among t-shirt buyers, they spend an average of \$45.00, with a standard deviation of \$20.10.

LINEAR TRANSFORMATIONS OF A RANDOM VARIABLE

If X is a random variable, then a linear transformation is given by $aX + b$, where a and b are some fixed numbers.

$$E(aX + b) = a \times E(X) + b \qquad\qquad SD(aX + b) = |a| \times SD(X)$$

3.5.6   Linear combinations of random variables

So far, we have thought of each variable as being a complete story in and of itself. Sometimes it is more appropriate to use a combination of variables. For instance, the amount of time a person spends commuting to work each week can be broken down into several daily commutes. Similarly, the total gain or loss in a stock portfolio is the sum of the gains and losses in its components.

EXAMPLE 3.90

John travels to work five days a week. We will use X_1 to represent his travel time on Monday, X_2 to represent his travel time on Tuesday, and so on. Write an equation using X_1, ..., X_5 that represents his travel time for the week, denoted by W.

His total weekly travel time is the sum of the five daily values:

$$W = X_1 + X_2 + X_3 + X_4 + X_5$$

Breaking the weekly travel time W into pieces provides a framework for understanding each source of randomness and is useful for modeling W.

EXAMPLE 3.91

It takes John an average of 18 minutes each day to commute to work. What would you expect his average commute time to be for the week?

We were told that the average (i.e. expected value) of the commute time is 18 minutes per day: $E(X_i) = 18$. To get the expected time for the sum of the five days, we can add up the expected time for each individual day:

$$
\begin{aligned}
E(W) &= E(X_1 + X_2 + X_3 + X_4 + X_5) \\
&= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 18 + 18 + 18 + 18 + 18 = 90 \text{ minutes}
\end{aligned}
$$

The expectation of the total time is equal to the sum of the expected individual times. More generally, the expectation of a sum of random variables is always the sum of the expectation for each random variable.

GUIDED PRACTICE 3.92

Ⓖ Elena is selling a TV at a cash auction and also intends to buy a toaster oven in the auction. If X represents the profit for selling the TV and Y represents the cost of the toaster oven, write an equation that represents the net change in Elena's cash.[72]

GUIDED PRACTICE 3.93

Ⓖ Based on past auctions, Elena figures she should expect to make about \$175 on the TV and pay about \$23 for the toaster oven. In total, how much should she expect to make or spend?[73]

GUIDED PRACTICE 3.94

Ⓖ Would you be surprised if John's weekly commute wasn't exactly 90 minutes or if Elena didn't make exactly \$152? Explain.[74]

Two important concepts concerning combinations of random variables have so far been introduced. First, a final value can sometimes be described as the sum of its parts in an equation. Second, intuition suggests that putting the individual average values into this equation gives the average value we would expect in total. This second point needs clarification – it is guaranteed to be true in what are called *linear combinations of random variables*.

A **linear combination** of two random variables X and Y is a fancy phrase to describe a combination

$$aX + bY$$

where a and b are some fixed and known numbers. For John's commute time, there were five random variables – one for each work day – and each random variable could be written as having a fixed coefficient of 1:

$$1X_1 + 1X_2 + 1X_3 + 1X_4 + 1X_5$$

For Elena's net gain or loss, the X random variable had a coefficient of $+1$ and the Y random variable had a coefficient of -1.

When considering the average of a linear combination of random variables, it is safe to plug in the mean of each random variable and then compute the final result. For a few examples of nonlinear combinations of random variables – cases where we cannot simply plug in the means – see the footnote.[75]

LINEAR COMBINATIONS OF RANDOM VARIABLES AND THE AVERAGE RESULT

If X and Y are random variables, then a linear combination of the random variables is given by $aX + bY$, where a and b are some fixed numbers. To compute the average value of a linear combination of random variables, plug in the average of each individual random variable and compute the result:

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

Recall that the expected value is the same as the mean, i.e. $E(X) = \mu_X$.

---

[72]She will make X dollars on the TV but spend Y dollars on the toaster oven: $X - Y$.
[73]$E(X - Y) = E(X) - E(Y) = 175 - 23 = \$152$. She should expect to make about \$152.
[74]No, since there is probably some variability. For example, the traffic will vary from one day to next, and auction prices will vary depending on the quality of the merchandise and the interest of the attendees.
[75]If X and Y are random variables, consider the following combinations: X^{1+Y}, $X \times Y$, X/Y. In such cases, plugging in the average value for each random variable and computing the result will not generally lead to an accurate average value for the end result.

EXAMPLE 3.95

Leonard has invested \$6000 in Google Inc. (stock ticker: GOOG) and \$2000 in Exxon Mobil Corp. (XOM). If X represents the change in Google's stock next month and Y represents the change in Exxon Mobil stock next month, write an equation that describes how much money will be made or lost in Leonard's stocks for the month.

For simplicity, we will suppose X and Y are not in percents but are in decimal form (e.g. if Google's stock increases 1%, then $X = 0.01$; or if it loses 1%, then $X = -0.01$). Then we can write an equation for Leonard's gain as

$$\$6000 \times X + \$2000 \times Y$$

If we plug in the change in the stock value for X and Y, this equation gives the change in value of Leonard's stock portfolio for the month. A positive value represents a gain, and a negative value represents a loss.

GUIDED PRACTICE 3.96

Suppose Google and Exxon Mobil stocks have recently been rising 2.1% and 0.4% per month, respectively. Compute the expected change in Leonard's stock portfolio for next month.[76]

GUIDED PRACTICE 3.97

You should have found that Leonard expects a positive gain in Guided Practice 3.96. However, would you be surprised if he actually had a loss this month?[77]

3.5.7   Variability in linear combinations of random variables

Quantifying the average outcome from a linear combination of random variables is helpful, but it is also important to have some sense of the uncertainty associated with the total outcome of that combination of random variables. The expected net gain or loss of Leonard's stock portfolio was considered in Guided Practice 3.96. However, there was no quantitative discussion of the volatility of this portfolio. For instance, while the average monthly gain might be about \$134 according to the data, that gain is not guaranteed. Figure 3.25 shows the monthly changes in a portfolio like Leonard's during the 36 months from 2009 to 2011. The gains and losses vary widely, and quantifying these fluctuations is important when investing in stocks.
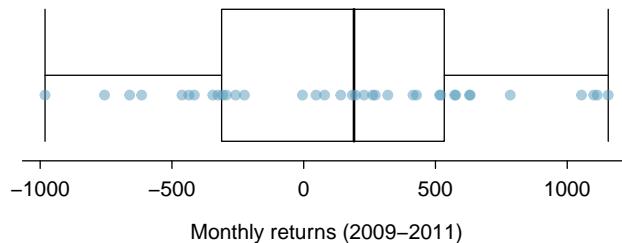
Figure 3.25: The change in a portfolio like Leonard's for the 36 months from 2009 to 2011, where \$6000 is in Google's stock and \$2000 is in Exxon Mobil's.

Just as we have done in many previous cases, we use the variance and standard deviation to describe the uncertainty associated with Leonard's monthly returns. To do so, the standard deviations and variances of each stock's monthly return will be useful, and these are shown in Figure 3.26. The stocks' returns are nearly independent.

---

[76] $E(\$6000 \times X + \$2000 \times Y) = \$6000 \times 0.021 + \$2000 \times 0.004 = \$134$.
[77] No. While stocks tend to rise over time, they are often volatile in the short term.

	Mean ($\bar{x}$)	Standard deviation (s)	Variance (s^2)
GOOG	0.0210	0.0849	0.0072
XOM	0.0038	0.0520	0.0027

Figure 3.26: The mean, standard deviation, and variance of the GOOG and XOM stocks. These statistics were estimated from historical stock data, so notation used for sample statistics has been used.

We want to describe the uncertainty of Leonard's monthly returns by finding the standard deviation of the return on his combined portfolio. First, we note that the variance of a sum has a nice property: the variance of a sum is the sum of the variances. That is, if X and Y are independent random variables:

$$Var(X + Y) = Var(X) + Var(Y)$$

Because the standard deviation is the square root of the variance, we can rewrite this equation using standard deviations:

$$(SD_{X+Y})^2 = (SD_X)^2 + (SD_Y)^2$$

This equation might remind you of a theorem from geometry: $c^2 = a^2 + b^2$. The equation for the standard deviation of the sum of two independent random variables looks analogous to the Pythagorean Theorem. Just as the Pythagorean Theorem only holds for right triangles, this equation only holds when X and Y are *independent*.[78]

STANDARD DEVIATION OF THE SUM AND DIFFERENCE OF RANDOM VARIABLES

If X and Y are *independent* random variables:

$$SD_{X+Y} = SD_{X-Y} = \sqrt{(SD_X)^2 + (SD_Y)^2}$$

Because $SD_Y = SD_{-Y}$, the standard deviation of the difference of two variables equals the standard deviation of the sum of two variables. This property holds for more than two variables as well. For example, if X, Y, and Z are independent random variables:

$$SD_{X+Y+Z} = SD_{X-Y-Z} = \sqrt{(SD_X)^2 + (SD_Y)^2 + (SD_Z)^2}$$

If we need the standard deviation of a linear combination of independent variables, such as $aX + bY$, we can consider aX and bY as two new variables. Recall that multiplying all of the values of variable by a positive constant multiplies the standard deviation by that constant. Thus, $SD_{aX} = a \times SD_X$ and $SD_{bY} = b \times SD_Y$. It follows that:

$$SD_{aX+bY} = \sqrt{(a \times SD_X)^2 + (b \times SD_Y)^2}$$

This equation can be used to compute the standard deviation of Leonard's monthly return. Recall that Leonard has $6,000 in Google stock and $2,000 in Exxon Mobil's stock. From Figure 3.26, the standard deviation of Google stock is 0.0849 and the standard deviation of Exxon Mobile stock is 0.0520.

$$
\begin{aligned}
SD_{6000X+2000Y} &= \sqrt{(6000 \times SD_X)^2 + (2000 \times SD_Y)^2} \\
&= \sqrt{(6000 \times 0.0849)^2 + (2000 \times .0520)^2} \\
&= \sqrt{270{,}304} = 520
\end{aligned}
$$

The standard deviation of the total is $520. While an average monthly return of $134 on an $8000 investment is nothing to scoff at, the monthly returns are so volatile that Leonard should not expect this income to be very stable.

---

[78]Another word for independent is orthogonal, meaning right angle! When X and Y are dependent, the equation for SD_{X+Y} becomes analogous to the law of cosines.

> **STANDARD DEVIATION OF LINEAR COMBINATIONS OF RANDOM VARIABLES**
>
> To find the standard deviation of a linear combination of random variables, we first consider aX and bY separately. We find the standard deviation of each, and then we apply the equation for the standard deviation of the sum of two variables:
>
> $$SD_{aX+bY} = \sqrt{(a \times SD_X)^2 + (b \times SD_Y)^2}$$
>
> This equation is valid as long as the random variables X and Y are *independent* of each other.

EXAMPLE 3.98

Suppose John's daily commute has a standard deviation of 4 minutes. What is the uncertainty in his total commute time for the week?

The expression for John's commute time is

$$X_1 + X_2 + X_3 + X_4 + X_5$$

Each coefficient is 1, so the standard deviation of the total weekly commute time is

$$
\begin{aligned}
SD &= \sqrt{(1 \times 4)^2 + (1 \times 4)^2 + (1 \times 4)^2 + (1 \times 4)^2 + (1 \times 4)^2} \\
&= \sqrt{5 \times (4)^2} \\
&= 8.94
\end{aligned}
$$

The standard deviation for John's weekly work commute time is about 9 minutes.

GUIDED PRACTICE 3.99

The computation in Example 3.98 relied on an important assumption: the commute time for each day is independent of the time on other days of that week. Do you think this is valid? Explain.[79]

GUIDED PRACTICE 3.100

Consider Elena's two auctions from Guided Practice 3.92 on page 183. Suppose these auctions are approximately independent and the variability in auction prices associated with the TV and toaster oven can be described using standard deviations of \$25 and \$8. Compute the standard deviation of Elena's net gain.[80]

Consider again Guided Practice 3.100. The negative coefficient for Y in the linear combination was eliminated when we squared the coefficients. This generally holds true: negatives in a linear combination will have no impact on the variability computed for a linear combination, but they do impact the expected value computations.

---

[79] One concern is whether traffic patterns tend to have a weekly cycle (e.g. Fridays may be worse than other days). If that is the case, and John drives, then the assumption is probably not reasonable. However, if John walks to work, then his commute is probably not affected by any weekly traffic cycle.

[80] The equation for Elena can be written as: $(1) \times X + (-1) \times Y$. To find the SD of this new variable we do:

$$SD_{(1) \times X + (-1) \times Y} = \sqrt{(1 \times SD_X)^2 + (-1 \times SD_Y)^2} = (1 \times 25)^2 + (-1 \times 8)^2 = 26.25$$

The SD is about \$26.25.

Section summary

- A **discrete probability distribution** can be summarized in a table that consists of all possible outcomes of a random variable and the probabilities of those outcomes. The outcomes must be disjoint, and the sum of the probabilities must equal 1.

- A probability distribution can be represented with a histogram and, like the distributions of data that we saw in Chapter 2, can be summarized by its **center**, **spread**, and **shape**.

- When given a probability distribution table, we can calculate the **mean** (expected value) and **standard deviation** of a random variable using the following formulas.

$$E(X) = \mu_X = \sum x_i \cdot P(x_i)$$
$$= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n)$$
$$Var(X) = \sigma_x^2 = \sum (x_i - \mu_X)^2 \cdot P(x_i)$$
$$SD(X) = \sigma_X = \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)}$$
$$= \sqrt{(x_1 - \mu_X)^2 \cdot P(x_1) + (x_2 - \mu_X)^2 \cdot P(x_2) + \cdots + (x_n - \mu_X)^2 \cdot P(x_n)}$$

We can think of $P(x_i)$ as the *weight*, and each term is weighted its appropriate amount.

- The **mean** of a probability distribution does not need to be a value in the distribution. It represents the average of many, many repetitions of a random process. The **standard deviation** represents the typical variation of the outcomes from the mean, when the random process is repeated over and over.

- **Linear transformations**. Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.

- **Combining random variables**. Let X and Y be random variables and let a and b be constants.

  - The expected value of the sum is the sum of the expected values.
    $$E(X + Y) = E(X) + E(Y)$$
    $$E(aX + bY) = a \times E(X) + b \times E(Y)$$

  - When X and Y are **independent**: The standard deviation of a sum or a difference is the square root of the sum of each standard deviation squared.
    $$SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$$
    $$SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$$
    $$SD(aX + bY) = \sqrt{(a \times SD(X))^2 + (b \times SD(Y))^2}$$

  The SD properties require that X and Y be independent. The expected value properties hold true whether or not X and Y are independent.

Exercises

3.37   College smokers. At a university, 13% of students smoke.

(a) Calculate the expected number of smokers in a random sample of 100 students from this university.

(b) The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

3.38   Ace of clubs wins. Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win $5. For any club, you win $10 plus an extra $20 for the ace of clubs.

(a) Create a probability model for the amount you win at this game. Also, find the expected winnings for a single game and the standard deviation of the winnings.

(b) What is the maximum amount you would be willing to pay to play this game? Explain your reasoning.

3.39   Hearts win. In a new card game, you start with a well-shuffled full deck and draw 3 cards without replacement. If you draw 3 hearts, you win $50. If you draw 3 black cards, you win $25. For any other draws, you win nothing.

(a) Create a probability model for the amount you win at this game, and find the expected winnings. Also compute the standard deviation of this distribution.

(b) If the game costs $5 to play, what would be the expected value and standard deviation of the net profit (or loss)? *(Hint: profit = winnings − cost; X − 5)*

(c) If the game costs $5 to play, should you play this game? Explain.

3.40   Is it worth it? Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs $2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins $3. For any ace, he wins $5, and he wins an *extra* $20 if he draws the ace of clubs.

(a) Create a probability model and find Andy's expected profit per game.

(b) Would you recommend this game to Andy as a good way to make money? Explain.

3.41   Portfolio return. A portfolio's value increases by 18% during a financial boom and by 9% during normal times. It decreases by 12% during a recession. What is the expected return on this portfolio if each scenario is equally likely?

3.42   Baggage fees. An airline charges the following baggage fees: $25 for the first bag and $35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

(a) Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

(b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

3.43   American roulette. The game of American roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money. Suppose you bet $1 on red. What's the expected value and standard deviation of your winnings?

3.44   European roulette. The game of European roulette involves spinning a wheel with 37 slots: 18 red, 18 black, and 1 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money.

(a) Suppose you play roulette and bet $3 on a single round. What is the expected value and standard deviation of your total winnings?

(b) Suppose you bet $1 in three different rounds. What is the expected value and standard deviation of your total winnings?

(c) How do your answers to parts (a) and (b) compare? What does this say about the riskiness of the two games?

3.6   Continuous distributions

So far we have looked only at cases where the random variable takes on integer values. What happens when we consider random variables that produce a continuous numerical variable, such as wait time for a bus? In this section, we introduce the concept of a continuous distribution. In the next chapter, you will encounter the most famous continuous distribution of all.[81]

Learning objectives

1. Understand the difference between a discrete random variable and a continuous random variable.

2. Recognize that when working with continuous probability distributions area represents probability and the total area under the curve must equal 1.

3.6.1   From histograms to continuous distributions

EXAMPLE 3.101

Figure 3.27 shows a few different hollow histograms of the variable `height` for 3 million US adults from the mid-90's.[82] How does changing the number of bins allow you to make different interpretations of the data?

Adding more bins provides greater detail. This sample is extremely large, which is why much smaller bins still work well. Usually we do not use so many bins with smaller sample sizes since small counts per bin mean the bin heights are very volatile.

EXAMPLE 3.102

What proportion of the sample is between 180 cm and 185 cm tall (about 5'11" to 6'1")?

We can add up the heights of the bins in the range 180 cm and 185 and divide by the sample size. For instance, this can be done with the two shaded bins shown in Figure 3.28. The two bins in this region have counts of 195,307 and 156,239 people, resulting in the following estimate of the probability:

$$\frac{195307 + 156239}{3,000,000} = 0.1172$$

This fraction is the same as the proportion of the histogram's area that falls in the range 180 to 185 cm.

---

[81]It's the normal distribution!
[82]This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.
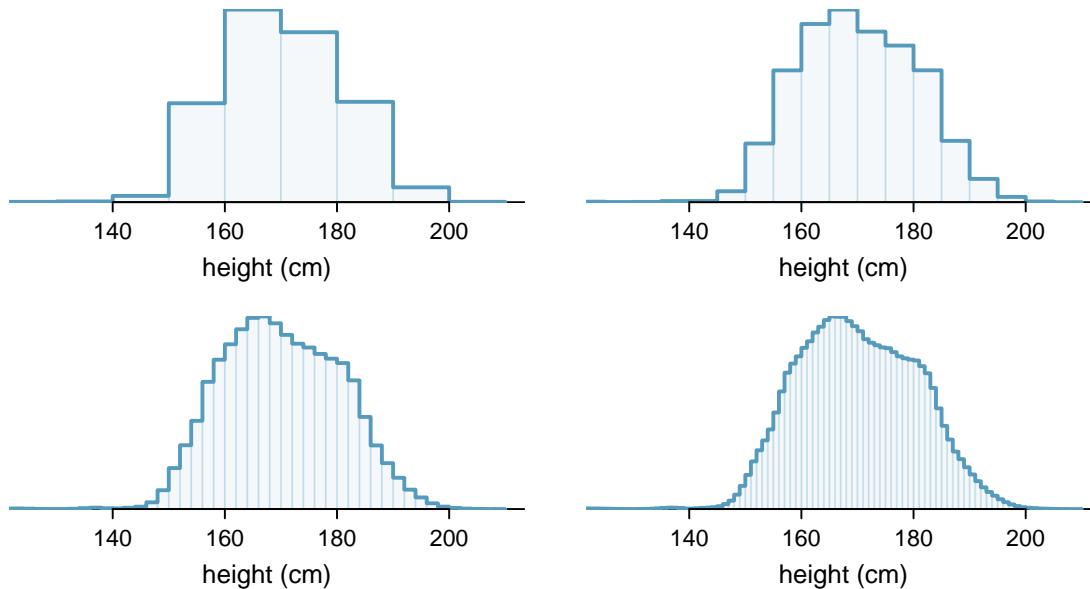
Figure 3.27: Four hollow histograms of US adults heights with varying bin widths.

Examine the transition from a boxy hollow histogram in the top-left of Figure 3.27 to the much smoother plot in the lower-right. In this last plot, the bins are so slim that the hollow histogram is starting to resemble a smooth curve. This suggests the population height as a *continuous* numerical variable might best be explained by a curve that represents the outline of extremely slim bins.

This smooth curve represents a **probability density function** (also called a **density** or **distribution**), and such a curve is shown in Figure 3.29 overlaid on a histogram of the sample. A density has a special property: the total area under the density's curve is 1.

3.6.2   Probabilities from continuous distributions

We computed the proportion of individuals with heights 180 to 185 cm in Example 3.102 as a fraction:

$$\frac{\text{number of people between 180 and 185}}{\text{total sample size}}$$

We found the number of people with heights between 180 and 185 cm by determining the fraction of the histogram's area in this region. Similarly, we can use the area in the shaded region under the curve to find a probability (with the help of a computer):

$$P(\texttt{height} \text{ between 180 and 185}) = \text{area between 180 and 185} = 0.1157$$

The probability that a randomly selected person is between 180 and 185 cm is 0.1157. This is very close to the estimate from Example 3.102: 0.1172.

GUIDED PRACTICE 3.103

Three US adults are randomly selected. The probability a single adult is between 180 and 185 cm is 0.1157.[83]

(a) What is the probability that all three are between 180 and 185 cm tall?

(b) What is the probability that none are between 180 and 185 cm?

---

[83]Brief answers: (a) $0.1157 \times 0.1157 \times 0.1157 = 0.0015$. (b) $(1 - 0.1157)^3 = 0.692$
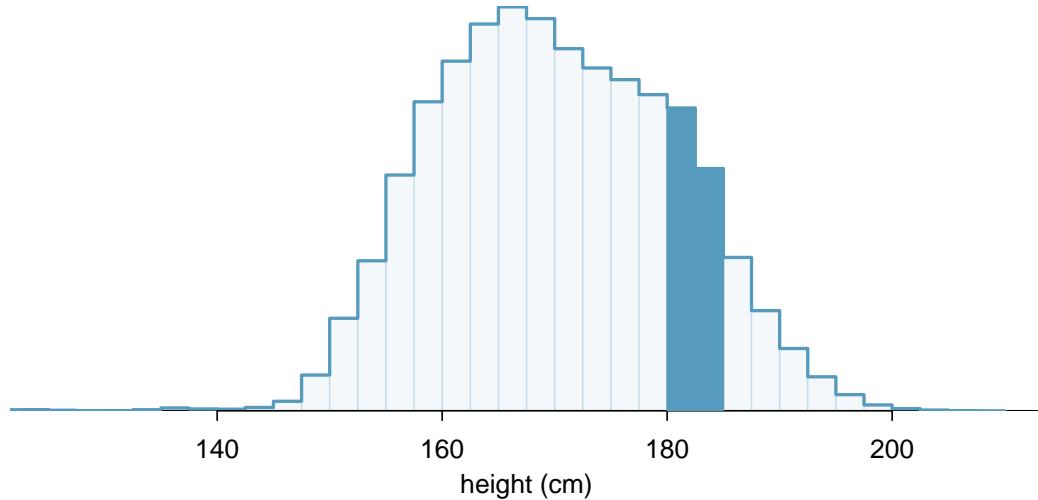
Figure 3.28: A histogram with bin sizes of 2.5 cm. The shaded region represents individuals with heights between 180 and 185 cm.
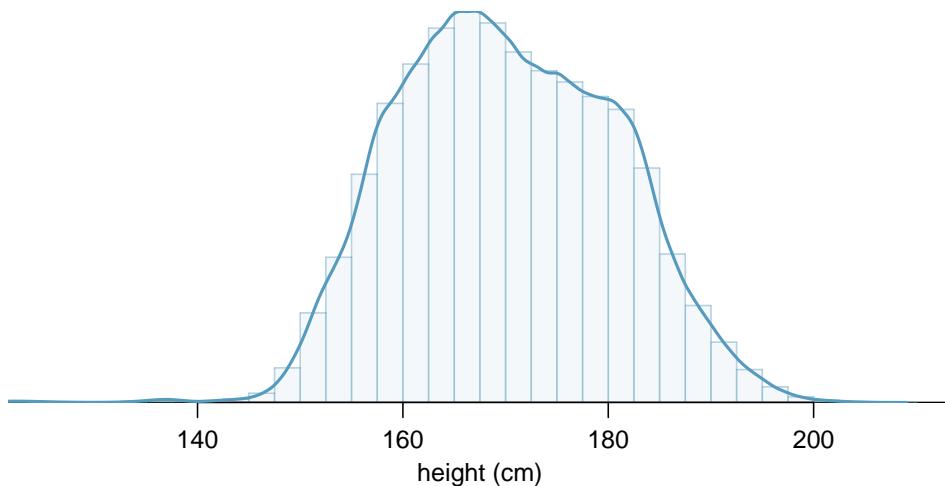
Figure 3.29: The continuous probability distribution of heights for US adults.

EXAMPLE 3.104

What is the probability that a randomly selected person is **exactly** 180 cm? Assume you can measure perfectly.

———

This probability is zero. A person might be close to 180 cm, but not exactly 180 cm tall. This also makes sense with the definition of probability as area; there is no area captured between 180 cm and 180 cm.

GUIDED PRACTICE 3.105

Suppose a person's height is rounded to the nearest centimeter. Is there a chance that a random person's **measured** height will be 180 cm?[84]

———————————————

[84]This has positive probability. Anyone between 179.5 cm and 180.5 cm will have a *measured* height of 180 cm. This is probably a more realistic scenario to encounter in practice versus Example 3.104.
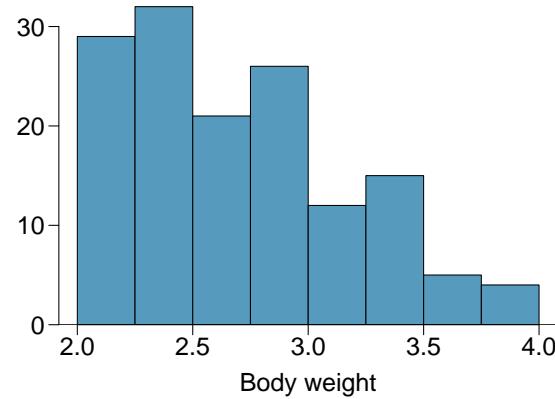
Section summary

- Histograms use bins with a specific width to display the distribution of a variable. When there is enough data and the data does not have gaps, as the bin width gets smaller and smaller, the histogram begins to resemble a smooth curve, or a **continuous distribution**.

- Continuous distributions are often used to approximate relative frequencies and probabilities. In a continuous distribution, the *area under the curve* corresponds to relative frequency or probability. The total area under a continuous probability distribution must equal 1.

- Because the area under the curve for a single point is zero, the probability of any specific value is zero. This implies that, for example, $P(X < 5) = P(X \leq 5)$ for a continuous probability distribution.

- Finding areas under curves is challenging; it is common to use distribution tables, calculators, or other technology to find such areas.

Exercises

3.45 Cat weights. The histogram shown below represents the weights (in kg) of 47 female and 97 male cats.[85]

(a) What fraction of these cats weigh less than 2.5 kg?

(b) What fraction of these cats weigh between 2.5 and 2.75 kg?

(c) What fraction of these cats weigh between 2.75 and 3.5 kg?

3.46 Income and gender. The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised of 59% males and 41% females.[86]

(a) Describe the distribution of total personal income.

(b) What is the probability that a randomly chosen US resident makes less than $50,000 per year?

(c) What is the probability that a randomly chosen US resident makes less than $50,000 per year and is female? Note any assumptions you make.

(d) The same data source indicates that 71.8% of females make less than $50,000 per year. Use this value to determine whether or not the assumption you made in part (c) is valid.

Income	Total
$1 to $9,999 or loss	2.2%
$10,000 to $14,999	4.7%
$15,000 to $24,999	15.8%
$25,000 to $34,999	18.3%
$35,000 to $49,999	21.2%
$50,000 to $64,999	13.9%
$65,000 to $74,999	5.8%
$75,000 to $99,999	8.4%
$100,000 or more	9.7%

[85]W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. www.stats.ox.ac.uk/pub/MASS4. New York: Springer, 2002.

[86]U.S. Census Bureau, 2005-2009 American Community Survey.

Chapter highlights

This chapter focused on understanding likelihood and chance variation, first by solving individual probability questions and then by investigating probability distributions.

The main probability techniques covered in this chapter are as follows:

- The **General Multiplication Rule** for **and** probabilities (intersection), along with the special case when events are **independent**.

- The **General Addition Rule** for **or** probabilities (union), along with the special case when events are **mutually exclusive**.

- The **Conditional Probability Rule**.

- Tree diagrams and **Bayes' Theorem** to solve more complex conditional problems.

- The **Binomial Formula** for finding the probability of exactly x successes in n independent trials.

- **Simulations** and the use of random digits to estimate probabilities.

Fundamental to all of these problems is understanding when events are independent and when they are mutually exclusive. Two events are **independent** when the outcome of one does not affect the outcome of the other, i.e. $P(A|B) = P(A)$. Two events are **mutually exclusive** when they cannot both happen together, i.e. $P(A \text{ and } B) = 0$.

Moving from solving individual probability questions to studying probability distributions helps us better understand chance processes and quantify expected chance variation.

- For a **discrete probability distribution**, the **sum** of the probabilities must equal 1. For a **continuous probability distribution**, the **area under the curve** represents a probability and the total area under the curve must equal 1.

- As with any distribution, one can calculate the mean and standard deviation of a probability distribution. In the context of a probability distribution, the **mean** and **standard deviation** describe the average and the typical deviation from the average, respectively, after many, many repetitions of the chance process.

- A probability distribution can be summarized by its **center** (mean, median), **spread** (SD, IQR), and **shape** (right skewed, left skewed, approximately symmetric).

- Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.

- The mean of the sum of two random variables equals the sum of the means. However, this is not true for standard deviations. Instead, when finding the standard deviation of a sum or difference of random variables, take the square root of the sum of each of the standard deviations squared.

The study of probability is useful for measuring uncertainty and assessing risk. In addition, probability serves as the foundation for inference, providing a framework for evaluating when an outcome falls outside of the range of what would be expected by chance alone.

Chapter exercises

3.47 Grade distributions. Each row in the table below is a proposed grade distribution for a class. Identify each as a valid or invalid probability distribution, and explain your reasoning.

	Grades				
	A	B	C	D	F
(a)	0.3	0.3	0.3	0.2	0.1
(b)	0	0	1	0	0
(c)	0.3	0.3	0.3	0	0
(d)	0.3	0.5	0.2	0.1	-0.1
(e)	0.2	0.4	0.2	0.1	0.1
(f)	0	-0.1	1.1	0	0

3.48 Health coverage, frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.[87]

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health	No	459	727	854	385	99	2,524
Coverage	Yes	4,198	6,245	4,821	1,634	578	17,476
	Total	4,657	6,972	5,675	2,019	677	20,000

(a) If we draw one individual at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

(b) If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

3.49 HIV in Swaziland. Swaziland has the highest HIV prevalence in the world: 25.9% of this country's population is infected with HIV.[88] The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

3.50 Twins. About 30% of human twins are identical, and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the probability that they are identical?

3.51 Cost of breakfast. Sally gets a cup of coffee and a muffin every day for breakfast from one of the many coffee shops in her neighborhood. She picks a coffee shop each morning at random and independently of previous days. The average price of a cup of coffee is $1.40 with a standard deviation of 30¢ ($0.30), the average price of a muffin is $2.50 with a standard deviation of 15¢, and the two prices are independent of each other.

(a) What is the mean and standard deviation of the amount she spends on breakfast daily?

(b) What is the mean and standard deviation of the amount she spends on breakfast weekly (7 days)?

---

[87]Office of Surveillance, Epidemiology, and Laboratory Services Behavioral Risk Factor Surveillance System, BRFSS 2010 Survey Data.

[88]Source: CIA Factbook, Country Comparison: HIV/AIDS - Adult Prevalence Rate.

3.52   Scooping ice cream. Ice cream usually comes in 1.5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances:

	mean	SD	variance
X	48	1	1
Y	2	0.25	0.0625

(a) An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?

(b) How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of $X - Y$. What is the standard deviation of the amount left in the box?

(c) Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

Chapter 4

Distributions of random variables

In this chapter, we discuss statistical distributions that frequently arise in the context of data analysis or statistical inference. We start with the normal distribution in the first section, and we find that it is useful for approximating many of the other distributions introduced in the sections that follow.

We also introduce the fundamental concept of a sampling distribution. The sampling distribution of a sample proportion and of a sample mean underlie the inference procedures encountered in next chapters.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

4.1   Normal distribution

What proportion of adults have systolic blood pressure above 140? What is the probability of getting more than 250 heads in 400 tosses of a fair coin? If the average weight of a piece of carry-on luggage is 11 pounds, what is the probability that 200 random carry on pieces will weigh more than 2500 pounds? If 55% of a population supports a certain candidate, what is the probability that she will have less than 50% support in a random sample of size 200?

There is one distribution that can help us answer all of these questions. Can you guess what it is? That's right – it's the normal distribution.

Learning objectives

1. Calculate and interpret a Z-score.

2. Understand that Z-scores are unitless (standard units) and are not affected by change of units.

3. Use the normal model to approximate a distribution where appropriate.

4. Find probabilities and percentiles using the normal approximation.

5. Find the value that corresponds to a given percentile when the distribution is approximately normal.

4.1.1   Normal distribution model

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**.[1] A normal curve is shown in Figure 4.1.
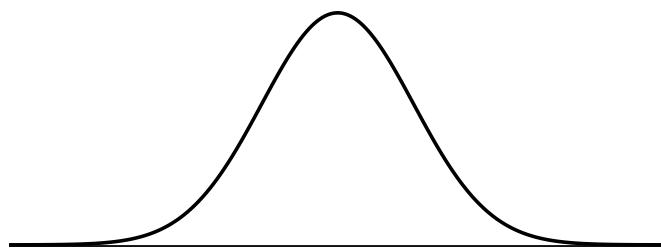
Figure 4.1: A normal curve.

---

[1] It is also introduced as the Gaussian distribution after Frederic Gauss, the first person to formalize its mathematical expression.

The **normal distribution** always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve. Figure 4.2 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure 4.3 shows these distributions on the same axis.
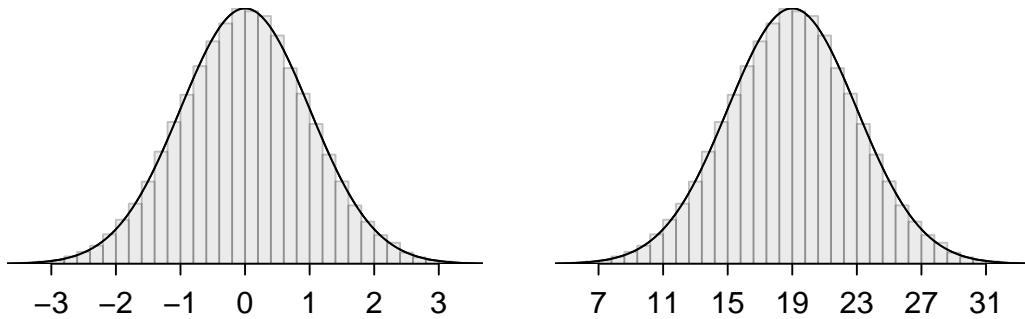
Figure 4.2: Both curves represent the normal distribution. However, they differ in their center and spread.
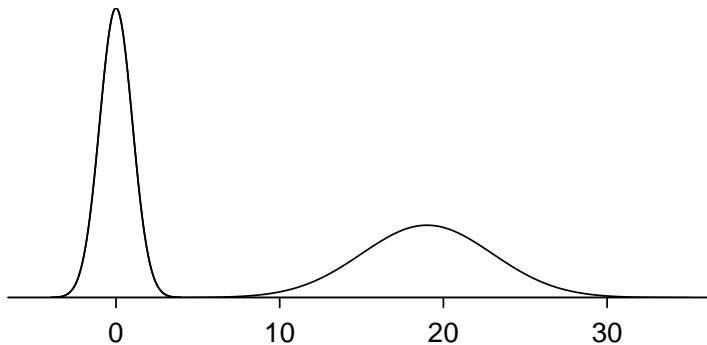
Figure 4.3: The normal distributions shown in Figure 4.2 but plotted together and on the same scale.

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**. The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**..

NORMAL DISTRIBUTION FACTS

Many variables are nearly normal, but none are exactly normal. The normal distribution, while never perfect, provides very close approximations for a variety of scenarios. We will use it to model data as well as probability distributions.

4.1.2 Standardizing with Z-scores

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

EXAMPLE 4.1

Figure 4.4 shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT: $1100 + 200 = 1300$. Tom is 0.5 standard deviations above the mean on the ACT: $21 + 0.5 \times 6 = 24$. In Figure 4.5, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

	SAT	ACT
Mean	1100	21
SD	200	6

Figure 4.4: Mean and standard deviation for the SAT and ACT.
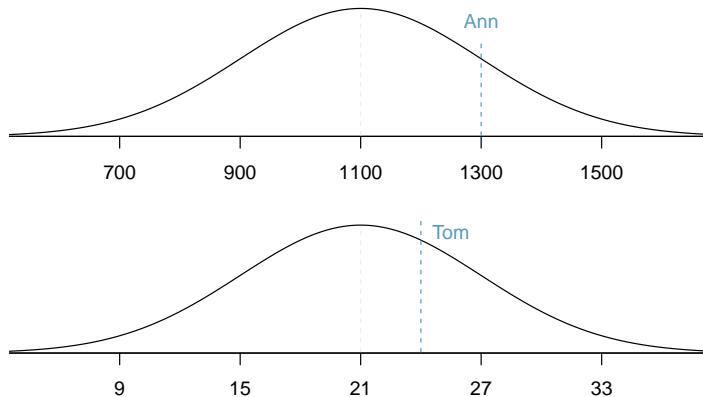
Figure 4.5: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

Example 4.1 used a standardization technique called a Z-score, a method most commonly employed for nearly normal observations but that may be used with any distribution. Recall from Chapter 2 that the **Z-score** of an observation is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its Z-score is 1. If it is 1.5 standard deviations *below* the mean, then its Z-score is -1.5. If x is an observation from a distribution with mean μ and standard deviation σ, we define the Z-score mathematically as

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

Using $\mu_{SAT} = 1100$, $\sigma_{SAT} = 200$, and $x_{\text{Ann}} = 1300$, we find Ann's Z-score:

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

GUIDED PRACTICE 4.2

Use Tom's ACT score, 24, along with the ACT mean and standard deviation to find his Z-score.[2]

---

[2] $Z_{Tom} = \frac{x_{\text{Tom}} - \mu_{\text{ACT}}}{\sigma_{\text{ACT}}} = \frac{24 - 21}{6} = 0.5$

4.1.3   Normal probability table

EXAMPLE 4.3

Ann from Example 4.1 earned a score of 1300 on her SAT with a corresponding $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than Ann. We shade the area representing those individuals in Figure 4.6. The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in Figure 4.6: 0.8413. In other words, Ann is in the 84th percentile of SAT takers.
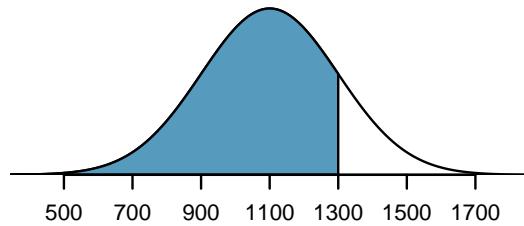
Figure 4.6: The normal model for SAT scores, shading the area of those individuals who scored below Ann.

We can use the normal model to find percentiles. A **normal probability table**, which lists Z-scores and corresponding percentiles, can be used to identify a percentile based on the Z-score (and vice versa). Statistical software can also be used.

A normal probability table is given in Appendix C.2 on page 521 and abbreviated in Figure 4.8. We use this table to identify the percentile corresponding to any particular Z-score. For instance, the percentile of $Z = 0.43$ is shown in row 0.4 and column 0.03 in Figure 4.8: 0.6664, or the 66.64^{th} percentile. Generally, we round Z to two decimals, identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation.
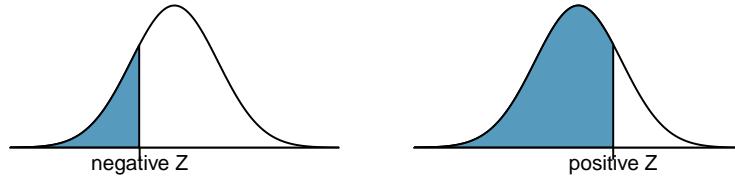
Figure 4.7: The area to the left of Z represents the percentile of the observation.

Z	Second decimal place of Z
	0.00	0.01	0.02	0.03	**0.04**	*0.05*	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	*0.6736*	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Figure 4.8: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.45$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

We can also find the Z-score associated with a percentile. For example, to identify Z for the 80th percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z-score for the 80th percentile by combining the row and column Z values: 0.84.

(G) **GUIDED PRACTICE 4.4**

Determine the proportion of SAT test takers who scored better than Ann on the SAT.[3]

---

[3]If 84% had lower scores than Ann, the proportion of people who had better scores must be 16%. (Generally ties are ignored when the normal model, or any other continuous distribution, is used.)
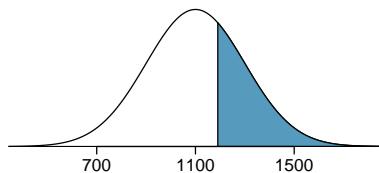
4.1.4   Normal probability examples

Cumulative SAT scores are approximated well by a normal model with mean 1100 and standard deviation 200.

EXAMPLE 4.5

What is the probability that a randomly selected SAT taker scores at least 1190 on the SAT?
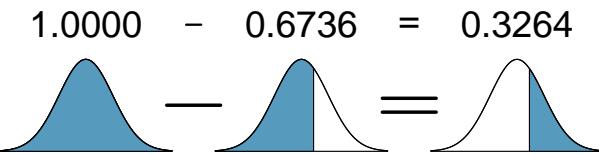
The probability that a randomly selected SAT taker scores at least 1190 on the SAT is equivalent to the proportion of all SAT takers that score at least 1190 on the SAT. First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the probability that a randomly selected score will be above 1190, so we shade this upper tail:

The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z-score of the cutoff value. With $\mu = 1100$, $\sigma = 200$, and the cutoff value $x = 1190$, the Z-score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

We look up the percentile of $Z = 0.45$ in the normal probability table shown in Figure 4.8 or in Appendix C.2 on page 521, which yields 0.6736. However, the percentile describes those who had a Z-score *lower* than 0.45. To find the area *above* $Z = 0.45$, we compute one minus the area of the lower tail:

The probability that a randomly selected score is at least 1190 on the SAT is 0.3264.

ALWAYS DRAW A PICTURE FIRST, AND FIND THE Z-SCORE SECOND

For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability.

After drawing a figure to represent the situation, identify the Z-score for the observation of interest.

GUIDED PRACTICE 4.6
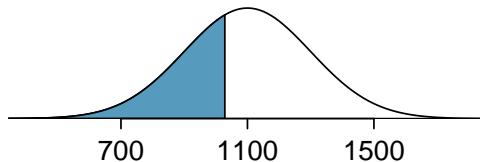
If the probability that a randomly selected score is at least 1190 is 0.3264, what is the probability that the score is less than 1190? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[4]

---

[4]We found the probability in Example 4.5: 0.6736. A picture for this exercise is represented by the shaded area below "0.6736" in Example 4.5.

EXAMPLE 4.7

Edward earned a 1030 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1030. These are the scores to the left of 1030.

Identifying the mean $\mu = 1100$, the standard deviation $\sigma = 200$, and the cutoff for the tail area $x = 1030$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Using the normal probability table, identify the row of -0.3 and column of 0.05, which corresponds to the probability 0.3632. Edward is at the 36th percentile.

GUIDED PRACTICE 4.8
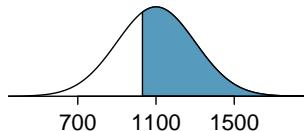
Use the results of Example 4.7 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[5]

AREAS TO THE RIGHT

The normal probability table in most books gives the area to the left. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

The last several problems have focused on finding the probability or percentile for a particular observation. It is also possible to identify the value corresponding to a particular percentile.

---
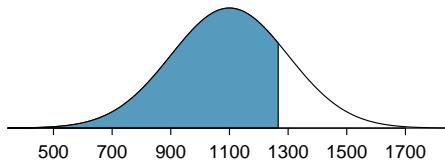
[5]If Edward did better than 36% of SAT takers, then about 64% must have done better than him.

EXAMPLE 4.9

Carlos believes he can get into his preferred college if he scores at least in the 80th percentile on the SAT. What score should he aim for?

Here, we are given a percentile rather than a Z-score, so we work backwards. As always, first draw the picture.

We want to find the observation that corresponds to the 80th percentile. First, we find the Z-score associated with the 80th percentile using the normal probability table. Looking at Figure 4.8., we look for the number closest to 0.80 *inside* the table. The closest number we find is 0.7995 (highlighted). 0.7995 falls on row 0.8 and column 0.04, therefore it corresponds to a Z-score of 0.84. In any normal distribution, a value with a Z-score of 0.84 will be at the 80th percentile. Once we have the Z-score, we work backwards to find x.

$$Z = \frac{x - \mu}{\sigma}$$
$$0.84 = \frac{x - 1100}{200}$$
$$0.84 \times 200 + 1100 = x$$
$$x = 1268$$

The 80th percentile on the SAT corresponds to a score of 1268.

GUIDED PRACTICE 4.10

Imani scored at the 72nd percentile on the SAT. What was her SAT score?[6]

IF THE DATA ARE NOT NEARLY NORMAL, DON'T USE A NORMAL TABLE

Before using the normal table, verify that the data or distribution is approximately normal. If it is not, the normal table will give incorrect results. Also, all answers based on normal approximations are approximations and are not exact.

---

[6]First, draw a picture! The closest percentile in the table to 0.72 is 0.7190, which corresponds to $Z = 0.58$. Next, set up the Z-score formula and solve for x: $0.58 = \frac{x-1100}{200} \rightarrow x = 1216$. Imani scored 1216.

4.1.5   Calculator: finding normal probabilities

TI-84: FINDING AREA UNDER THE NORMAL CURVE

Use 2ND VARS, normalcdf to find an area/proportion/probability between two Z-scores or to the left or right of a Z-score.

1. Choose 2ND VARS (i.e. DISTR).

2. Choose 2:normalcdf.

3. Enter the lower (left) Z-score and the upper (right) Z-score.

   - If finding just a lower tail area, set lower to -5.
   - If finding just an upper tail area, set upper to 5.

4. Leave μ as 0 and σ as 1.

5. Down arrow, choose Paste, and hit ENTER.

TI-83: Do steps 1-2, then enter the lower bound and upper bound separated by a comma, e.g. normalcdf(2, 5), and hit ENTER.
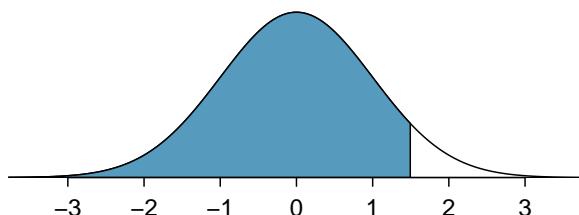
CASIO FX-9750GII: FINDING AREA UNDER THE NORMAL CURVE

1. Navigate to STAT (MENU, then hit 2).

2. Select DIST (F5), then NORM (F1), and then Ncd (F2).

3. If needed, set Data to Variable (Var option, which is F2).

4. Enter the Lower Z-score and the Upper Z-score. Set σ to 1 and μ to 0.

   - If finding just a lower tail area, set Lower to -5.
   - For an upper tail area, set Upper to 5.

5. Hit EXE, which will return the area probability (p) along with the Z-scores for the lower and upper bounds.

EXAMPLE 4.11

Use a calculator to determine what percentile corresponds to a Z-score of 1.5.

Always first sketch a graph:[7]

To find an area under the normal curve using a calculator, first identify a lower bound and an upper bound. Theoretically, we want all of the area to the left of 1.5, so the left endpoint should be -∞. However, the area under the curve is nearly negligible when Z is smaller than -4, so we will use -5 as the lower bound when not given a lower bound (any other negative number smaller than -5 will also work). Using a lower bound of -5 and an upper bound of 1.5, we get $P(Z < 1.5) = 0.933$.

---

[7]normalcdf gives the result without drawing the graph. To draw the graph, do 2nd VARS, DRAW, 1:ShadeNorm. However, beware of errors caused by other plots that might interfere with this plot.

GUIDED PRACTICE 4.12

Find the area under the normal curve to right of $Z = 2$. [8]

GUIDED PRACTICE 4.13

Find the area under the normal curve between -1.5 and 1.5. [9]

▶ TI-84: FIND A Z-SCORE THAT CORRESPONDS TO A PERCENTILE

Use 2ND VARS, invNorm to find the Z-score that corresponds to a given percentile.

1. Choose 2ND VARS (i.e. DISTR).
2. Choose 3:invNorm.
3. Let Area be the percentile as a decimal (the area to the left of desired Z-score).
4. Leave μ as 0 and σ as 1.
5. Down arrow, choose Paste, and hit ENTER.

TI-83: Do steps 1-2, then enter the percentile as a decimal, e.g. invNorm(.40), then hit ENTER.

▶ CASIO FX-9750GII: FIND A Z-SCORE THAT CORRESPONDS TO A PERCENTILE

1. Navigate to STAT (MENU, then hit 2).
2. Select DIST (F5), then NORM (F1), and then InvN (F3).
3. If needed, set Data to Variable (Var option, which is F2).
4. Decide which tail area to use (Tail), the tail area (Area), and then enter the σ and μ values.
5. Hit EXE.

EXAMPLE 4.14

Use a calculator to find the Z-score that corresponds to the 40th percentile.

———————

Letting Area be 0.40, a calculator gives -0.253. This means that $Z = -0.253$ corresponds to the 40th percentile, that is, $P(Z < -0.253) = 0.40$.

GUIDED PRACTICE 4.15

Find the Z-score such that 20 percent of the area is to the right of that Z-score. [10]

---

[8]Now we want to shade to the right. Therefore our lower bound will be 2 and the upper bound will be +5 (or a number bigger than 5) to get $P(Z > 2) = 0.023$.

[9]Here we are given both the lower and the upper bound. Lower bound is -1.5 and upper bound is 1.5. The area under the normal curve between -1.5 and 1.5 = $P(-1.5 < Z < 1.5) = 0.866$.

[10]If 20% of the area is the right, then 80% of the area is to the left. Letting area be 0.80, we get $Z = 0.841$.

EXAMPLE 4.16

In a large study of birth weight of newborns, the weights of 23,419 newborn boys were recorded.[11] The distribution of weights was approximately normal with a mean of 7.44 lbs (3376 grams) and a standard deviation of 1.33 lbs (603 grams). The government classifies a newborn as having low birth weight if the weight is less than 5.5 pounds. What percent of these newborns had a low birth weight?

We find an area under the normal curve between -5 (or a number smaller than -5, e.g. -10) and a Z-score that we will calculate. There is no need to write calculator commands in a solution. Instead, continue to use standard statistical notation.

$$Z = \frac{5.5 - 7.44}{1.33}$$
$$= -1.49$$
$$P(Z < -1.49) = 0.068$$

Approximately 6.8% of the newborns were of low birth weight.

GUIDED PRACTICE 4.17

Approximately what percent of these babies weighed greater than 10 pounds?[12]

GUIDED PRACTICE 4.18

Approximately *how many* of these newborns weighed greater than 10 pounds?[13]

GUIDED PRACTICE 4.19

How much would a newborn have to weigh in order to be at the 90th percentile among this group?[14]

---

[11]www.biomedcentral.com/1471-2393/8/5

[12]$Z = \frac{10-7.44}{1.33} = 1.925$. Using a lower bound of 2 and an upper bound of 5, we get $P(Z > 1.925) = 0.027$. Approximately 2.7% of the newborns weighed over 10 pounds.

[13]Approximately 2.7% of the newborns weighed over 10 pounds. Because there were 23,419 of them, about $0.027 \times 23419 \approx 632$ weighed greater than 10 pounds.

[14]Because we have the percentile, this is the inverse problem. To get the Z-score, use the inverse normal option with 0.90 to get $Z = 1.28$. Then solve for x in $1.28 = \frac{x-7.44}{1.33}$ to get $x = 9.15$. To be at the 90th percentile among this group, a newborn would have to weigh 9.15 pounds.

4.1.6   68-95-99.7 rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z table.
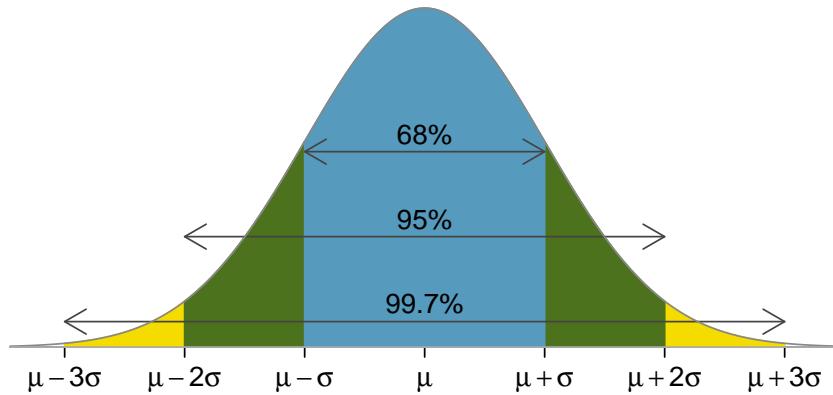
Figure 4.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

GUIDED PRACTICE 4.20

Use the Z table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[15]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-15,000. For 5 and 6 standard deviations, it is about 1-in-2 million and 1-in-500 million, respectively.

GUIDED PRACTICE 4.21

SAT scores closely follow the normal model with mean $\mu = 1100$ and standard deviation $\sigma = 200$. (a) About what percent of test takers score 700 to 1500? (b) What percent score between 1100 and 1500?[16]

---

[15]First draw the pictures. To find the area between $Z = -1$ and $Z = 1$, use the normal probability table to determine the areas below $Z = -1$ and above $Z = 1$. Next verify the area between $Z = -1$ and $Z = 1$ is about 0.68. Repeat this for $Z = -2$ to $Z = 2$ and also for $Z = -3$ to $Z = 3$.

[16](a) 700 and 1500 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 700 and 1500. (b) Since the normal model is symmetric, then half of the test takers from part (a) ($\frac{95\%}{2} = 47.5\%$ of all test takers) will score 700 to 1500 while 47.5% score between 1100 and 1500.

4.1.7 Evaluating the normal approximation

It is important to remember normality is always an approximation. Testing the appropriateness of the normal assumption is a key step in many data analyses.

The distribution of heights of US males is well approximated by the normal model. We are interested in proceeding under the assumption that the data are normally distributed, but first we must check to see if this is reasonable.

There are two visual methods for checking the assumption of normality that can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 4.10. The sample mean $\bar{x}$ and standard deviation s are used as the parameters of the best fitting normal curve. The closer this curve fits the histogram, the more reasonable the normal model assumption. Another more common method is examining a **normal probability plot**,[17] shown in the right panel of Figure 4.10. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.
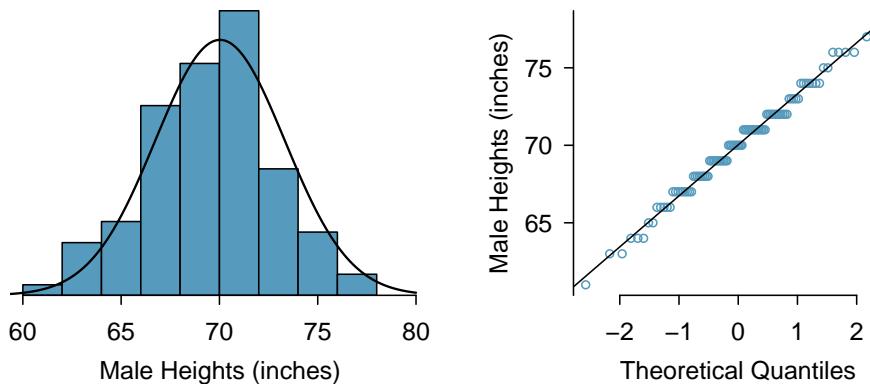
Figure 4.10: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

Three data sets of 40, 100, and 400 samples were simulated from a normal distribution, and the histograms and normal probability plots of the data sets are shown in Figure 4.11. These will provide a benchmark for what to look for in plots of real data.

The left panels show the histogram (top) and normal probability plot (bottom) for the simulated data set with 40 observations. The data set is too small to really see clear structure in the histogram. The normal probability plot also reflects this, where there are some deviations from the line. However, these deviations are not strong.

The middle panels show diagnostic plots for the data set with 100 simulated observations. The histogram shows more normality and the normal probability plot shows a better fit. While there is one observation that deviates noticeably from the line, it is not particularly extreme.

The data set with 400 observations has a histogram that greatly resembles the normal distribution, while the normal probability plot is nearly a perfect straight line. Again in the normal probability plot there is one observation (the largest) that deviates slightly from the line. If that observation had deviated 3 times further from the line, it would be of much greater concern in a real data set. Apparent outliers can occur in normally distributed data but they are rare.

Notice the histograms look more normal as the sample size increases, and the normal probability plot becomes straighter and more stable.

---

[17]Also commonly called a **quantile-quantile plot**.
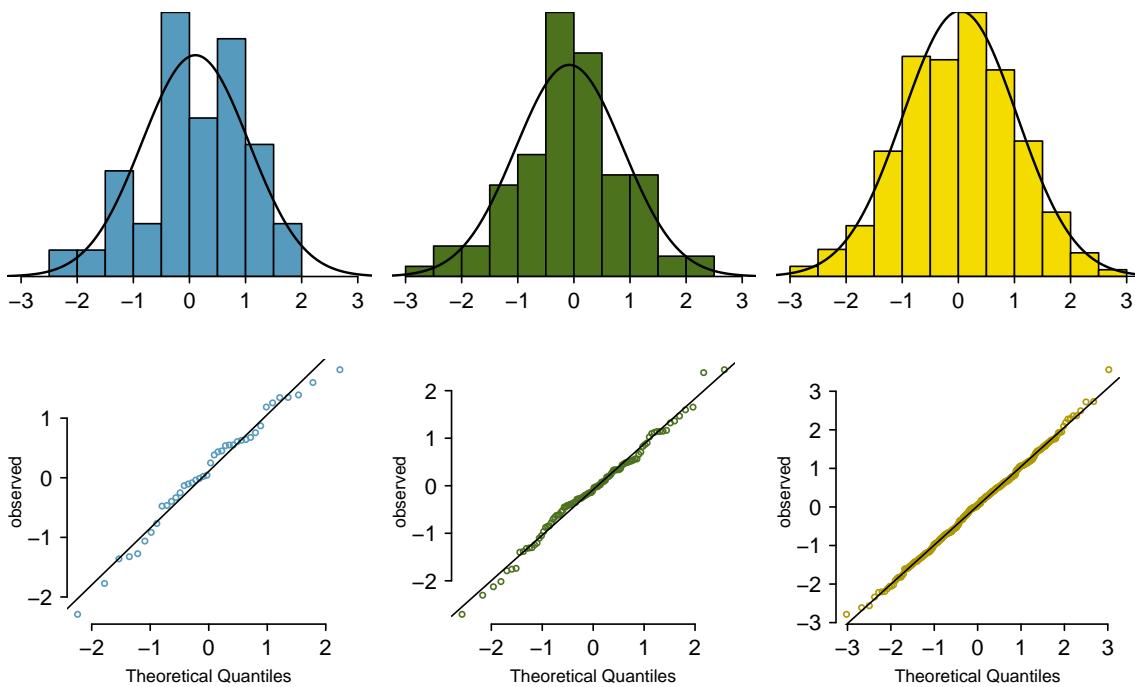
Figure 4.11: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

EXAMPLE 4.22

Consider all NBA players from the 2018-2019 season presented in Figure 4.12.[18]   Based on the graphs, are NBA player heights normally distributed?

We first create a histogram and normal probability plot of the NBA player heights. The histogram in the left panel is slightly left skewed, which contrasts with the symmetric normal distribution. The points in the normal probability plot do not appear to closely follow a straight line but show what appears to be a "wave". We can compare these characteristics to the sample of 400 normally distributed observations in Example 4.1.7 and see that they represent much stronger deviations from the normal model. NBA player heights do not appear to come from a normal distribution.

EXAMPLE 4.23

Consider the poker winnings of an individual over 50 days. A histogram and normal probability plot of these data are shown in Figure 4.13. Based on the graphs, can we approximate poker winnings by a normal distribution?

The data are very strongly right skewed in the histogram, which corresponds to the very strong deviations on the upper right component of the normal probability plot. If we compare these results to the sample of 40 normal observations in Example 4.1.7, it is apparent that these data show very strong deviations from the normal model.

---

[18]These data were collected from www.nba.com.
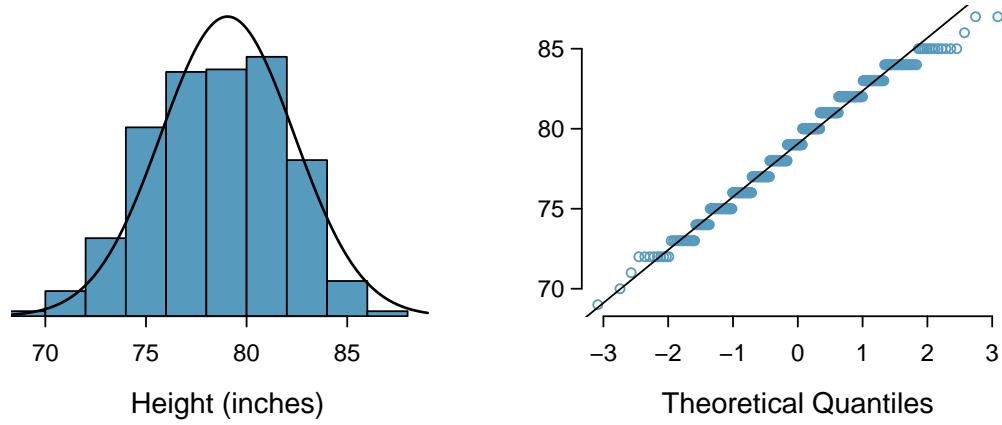
Figure 4.12: Histogram and normal probability plot for the NBA heights from the 2018-2019 season.
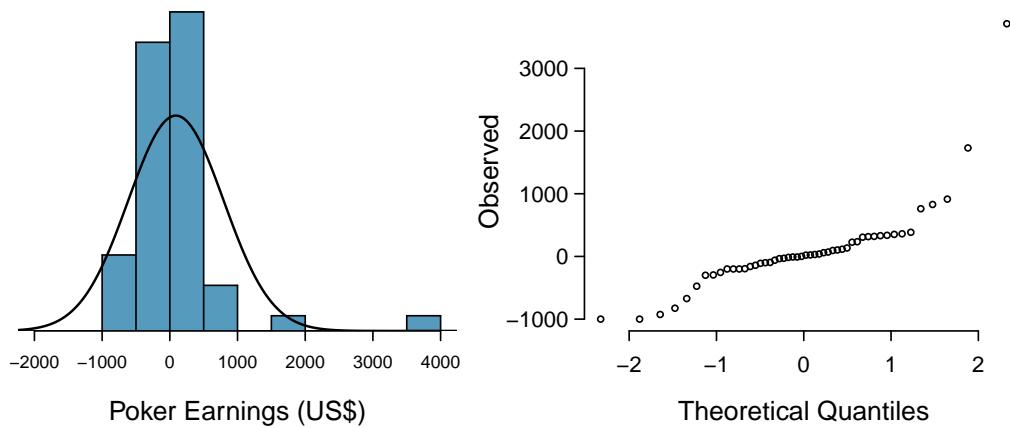
Figure 4.13: A histogram of poker data with the best fitting normal plot and a normal probability plot.

GUIDED PRACTICE 4.24

Determine which data sets represented in Figure 4.14 plausibly come from a nearly normal distribution. Are you confident in all of your conclusions? There are 100 (top left), 50 (top right), 500 (bottom left), and 15 points (bottom right) in the four plots.[19]
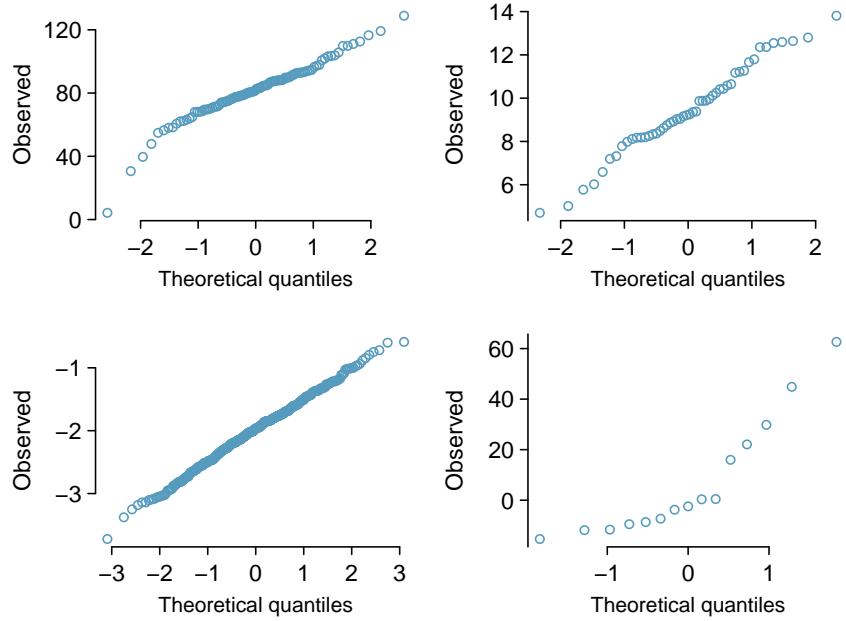
Figure 4.14: Four normal probability plots for Guided Practice 4.24.

GUIDED PRACTICE 4.25

Figure 4.15 shows normal probability plots for two distributions that are skewed. One distribution is skewed to the low end (left skewed) and the other to the high end (right skewed). Which is which?[20]
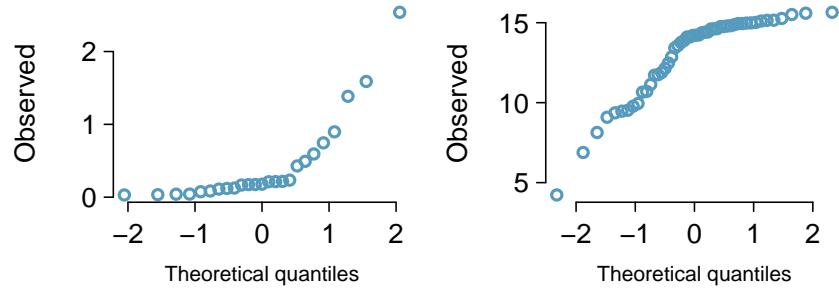
Figure 4.15: Normal probability plots for Guided Practice 4.25.

---

[19]Answers may vary a little. The top-left plot shows some deviations in the smallest values in the data set; specifically, the left tail of the data set has some outliers we should be wary of. The top-right and bottom-left plots do not show any obvious or extreme deviations from the lines for their respective sample sizes, so a normal model would be reasonable for these data sets. The bottom-right plot has a consistent curvature that suggests it is not from the normal distribution. If we examine just the vertical coordinates of these observations, we see that there is a lot of data between -20 and 0, and then about five observations scattered between 0 and 70. This describes a distribution that has a strong right skew.

[20]Examine where the points fall along the vertical axis. In the first plot, most points are near the low end with fewer observations scattered along the high end; this describes a distribution that is skewed to the high end. The second plot shows the opposite features, and this distribution is skewed to the low end.

4.1.8   Normal approximation for sums of random variables

We have seen that many distributions are approximately normal. The sum and the difference of normally distributed variables is also normal. While we cannot prove this here, the usefulness of it is seen in the following example.

EXAMPLE 4.26

Three friends are playing a cooperative video game in which they have to complete a puzzle as fast as possible. Assume that the individual times of the 3 friends are independent of each other. The individual times of the friends in similar puzzles are approximately normally distributed with the following means and standard deviations.

	Mean	SD
Friend 1	5.6	0.11
Friend 2	5.8	0.13
Friend 3	6.1	0.12

To advance to the next level of the game, the friends' total time must not exceed 17.1 minutes. What is the probability that they will advance to the next level?

Because each friend's time is approximately normally distributed, *the sum of their times is also approximately normally distributed*. We will do a normal approximation, but first we need to find the mean and standard deviation of the *sum*. We learned how to do this in Section 3.5.

Let the three friends be labeled X, Y, Z. We want $P(X + Y + Z < 17.1)$. The mean and standard deviation of the sum of X, Y, and Z is given by:

$$\mu_{sum} = E(X + Y + Z) \qquad\qquad \sigma_{sum} = \sqrt{(SD_X)^2 + (SD_Y)^2 + (SD_Z)^2}$$

$$= E(X) + E(Y) + E(Z) \qquad\qquad = \sqrt{(0.11)^2 + (0.13)^2 + (0.12)^2}$$

$$= 5.6 + 5.8 + 6.1 \qquad\qquad = 0.208$$

$$= 17.5$$

Now we can find the Z-score.

$$Z = \frac{x_{sum} - \mu_{sum}}{\sigma_{sum}}$$

$$= \frac{17.1 - 17.5}{0.208}$$

$$= -1.92$$

Finally, we want the probability that the sum is less than 17.5, so we shade the area to the left of $Z = -1.92$. Using the normal table or a calculator, we get

$$P(Z < -1.92) = 0.027$$

There is a 2.7% chance that the friends will advance to the next level.

GUIDED PRACTICE 4.27

What is the probability that Friend 2 will complete the puzzle with a faster time than Friend 1? Hint: find $P(Y < X)$, or $P(Y - X < 0)$.[21]

---

[21]First find the mean and standard deviation of $Y - X$. The mean of $Y - X$ is $\mu_{Y-X} = 5.8 - 5.6 = 0.2$. The standard deviation is $SD_{Y-X} = \sqrt{(0.13)^2 + (0.11)^2} = 0.170$. Then $Z = \frac{0 - 0.2}{0.170} = -1.18$ and $P(Z < -1.18) = .119$. There is an 11.9% chance that Friend 2 will complete the puzzle with a faster time than Friend 1.

Section summary

- A **Z-score** represents the number of standard deviations a value in a data set is above or below the mean. To calculate a Z-score use: $Z = \frac{x - \text{mean}}{SD}$.

- *Z-scores do not depend on units.* When looking at distributions with different units or different standard deviations, Z-scores are useful for comparing how far values are away from the mean (relative to the distribution of the data).

- The **normal distribution** is the most commonly used distribution in Statistics. Many distribution are approximately normal, but none are exactly normal.

- The empirical rule (68-95-99.7 Rule) comes from the normal distribution. The closer a distribution is to normal, the better this rule will hold.

- It is often useful to use the standard normal distribution, which has mean 0 and SD 1, to approximate a discrete histogram. There are two common types of **normal approximation problems**, and for each a key step is to find a Z-score.

    A: *Find the percent or probability of a value greater/less than a given x-value.*
        1. Verify that the distribution of interest is approximately normal.
        2. Calculate the Z-score. Use the provided population mean and SD to standardize the given x-value.
        3. Use a calculator function (e.g. `normcdf` on a TI) or a normal table to find the area under the normal curve to the right/left of this Z-score; this is the *estimate* for the percent/probability.

    B: *Find the x-value that corresponds to a given percentile.*
        1. Verify that the distribution of interest is approximately normal.
        2. Find the Z-score that corresponds to the given percentile (using, for example, `invNorm` on a TI).
        3. Use the Z-score along with the given mean and SD to solve for the x-value.

- Because the sum or difference of two normally distributed variables is itself a normally distributed variable, the normal approximation is also used in the following type of problem.

    Find the probability that a sum $X + Y$ or a difference $X - Y$ is greater/less than some value.
        1. Verify that the distribution of X and the distribution of Y are approximately normal.
        2. Find the mean of the sum or difference. Recall: the mean of a sum is the sum of the means. The mean of a difference is the difference of the means.
           Find the SD of the sum or difference using:
           $SD(X + Y) = SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$.
        3. Calculate the Z-score. Use the calculated mean and SD to standardize the given sum or difference.
        4. Find the appropriate area under the normal curve.

Exercises

4.1 Area under the curve, Part I. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$ (b) $Z > 1.48$ (c) $-0.4 < Z < 1.5$ (d) $|Z| > 2$

4.2 Area under the curve, Part II. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z > -1.13$ (b) $Z < 0.18$ (c) $Z > 8$ (d) $|Z| < 0.5$

4.3 GRE scores, Part I. Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

(a) What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z-scores.
(b) What do these Z-scores tell you?
(c) Relative to others, which section did she do better on?
(d) Find her percentile scores for the two exams.
(e) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?
(f) Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.
(g) If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

4.4 Triathlon times, Part I. In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the *Men, Ages 30 - 34* group while Mary competed in the *Women, Ages 25 - 29* group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the *Men, Ages 30 - 34* group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the *Women, Ages 25 - 29* group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

(a) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?
(b) Did Leo or Mary rank better in their respective groups? Explain your reasoning.
(c) What percent of the triathletes did Leo finish faster than in his group?
(d) What percent of the triathletes did Mary finish faster than in her group?
(e) If the distributions of finishing times are not nearly normal, would your answers to parts (a) - (d) change? Explain your reasoning.

4.5 GRE scores, Part II. In Exercise 4.3 we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Use this information to compute each of the following:

(a) The score of a student who scored in the 80th percentile on the Quantitative Reasoning section.
(b) The score of a student who scored worse than 70% of the test takers in the Verbal Reasoning section.

4.6   Triathlon times, Part II. In Exercise 4.4 we saw two distributions for triathlon times: $N(\mu = 4313, \sigma = 583)$ for *Men, Ages 30 - 34* and $N(\mu = 5261, \sigma = 807)$ for the *Women, Ages 25 - 29* group. Times are listed in seconds. Use this information to compute each of the following:

(a) The cutoff time for the fastest 5% of athletes in the men's group, i.e. those who took the shortest 5% of time to finish.

(b) The cutoff time for the slowest 10% of athletes in the women's group.

4.7   LA weather, Part I. The average daily high temperature in June in LA is 77°F with a standard deviation of 5°F. Suppose that the temperatures in June closely follow a normal distribution.

(a) What is the probability of observing an 83°F temperature or higher in LA during a randomly chosen day in June?

(b) How cool are the coldest 10% of the days (days with lowest average high temperature) during June in LA?

4.8   CAPM. The Capital Asset Pricing Model (CAPM) is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 14.7% (i.e. an average gain of 14.7%) with a standard deviation of 33%. A return of 0% means the value of the portfolio doesn't change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money.

(a) What percent of years does this portfolio lose money, i.e. have a return less than 0%?

(b) What is the cutoff for the highest 15% of annual returns with this portfolio?

4.9   LA weather, Part II. Exercise 4.7 states that average daily high temperature in June in LA is 77°F with a standard deviation of 5°F, and it can be assumed that they to follow a normal distribution. We use the following equation to convert °F (Fahrenheit) to °C (Celsius):

$$C = (F - 32) \times \frac{5}{9}.$$

(a) What is the probability of observing a 28°C (which roughly corresponds to 83°F) temperature or higher in June in LA? Calculate using the °C model from part (a).

(b) Did you get the same answer or different answers in part (b) of this question and part (a) of Exercise 4.7? Are you surprised? Explain.

(c) Estimate the IQR of the temperatures (in °C) in June in LA.

4.10   Find the SD. Cholesterol levels for women aged 20 to 34 follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dl). Women with cholesterol levels above 220 mg/dl are considered to have high cholesterol and about 18.5% of women fall into this category. Find the standard deviation of this distribution.

4.11   Scores on stats final, Part I. Below are final exam scores of 20 Introductory Statistics students.

$$\overset{\scriptstyle 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13 \quad 14 \quad 15 \quad 16 \quad 17 \quad 18 \quad 19 \quad 20}{57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94}$$

The mean score is 77.7 points. with a standard deviation of 8.44 points. Use this information to determine if the scores approximately follow the 68-95-99.7% Rule.

4.12   Heights of female college students, Part I. Below are heights of 25 female college students.

$$\overset{\scriptstyle 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13 \quad 14 \quad 15 \quad 16 \quad 17 \quad 18 \quad 19 \quad 20 \quad 21 \quad 22 \quad 23 \quad 24 \quad 25}{54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61, 61, 62, 62, 63, 63, 63, 64, 65, 65, 67, 67, 69, 73}$$

The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

4.13   Lemonade at The Cafe. Drink pitchers at The Cafe are intended to hold about 64 ounces of lemonade and glasses hold about 12 ounces. However, when the pitchers are filled by a server, they do not always fill it with exactly 64 ounces. There is some variability. Similarly, when they pour out some of the lemonade, they do not pour exactly 12 ounces. The amount of lemonade in a pitcher is normally distributed with mean 64 ounces and standard deviation 1.732 ounces. The amount of lemonade in a glass is normally distributed with mean 12 ounces and standard deviation 1 ounce.

(a) How much lemonade would you expect to be left in a pitcher after pouring one glass of lemonade?

(b) What is the standard deviation of the amount left in a pitcher after pouring one glass of lemonade?

(c) What is the probability that more than 50 ounces of lemonade is left in a pitcher after pouring one glass of lemonade?

4.14   Spray paint, Part I. Suppose the area that can be painted using a single can of spray paint is slightly variable and follows a nearly normal distribution with a mean of 25 square feet and a standard deviation of 3 square feet. Suppose also that you buy three cans of spray paint.

(a) How much area would you expect to cover with these three cans of spray paint?

(b) What is the standard deviation of the area you expect to cover with these three cans of spray paint?

(c) The area you wanted to cover is 80 square feet. What is the probability that you will be able to cover this entire area with these three cans of spray paint?

4.15   GRE scores, Part III. In Exercises 4.3 and 4.5 we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Suppose performance on these two sections is independent. Use this information to compute each of the following:

(a) The probability of a combined (verbal + quantitative) score above 320.

(b) The score of a student who scored better than 90% of the test takers overall.

4.16   Betting on dinner, Part I. Suppose a restaurant is running a promotion where prices of menu items are random following some underlying distribution. If you're lucky, you can get a basket of fries for $3, or if you're not so lucky you might end up having to pay $10 for the same menu item. The price of basket of fries is drawn from a normal distribution with mean $6 and standard deviation of $2. The price of a fountain drink is drawn from a normal distribution with mean $3 and standard deviation of $1. What is the probability that you pay more than $10 for a dinner consisting of a basket of fries and a fountain drink?

4.2   Sampling distribution of a sample mean

If bags of chips are produced with an average weight of 15 oz and a standard deviation of 0.1 oz, what is the probability that the average weight of 30 bags will be within 0.1 oz of the mean? The answer is not 68%! To answer this question we must visualize and understand what is called the *sampling distribution* of a sample mean.

Learning objectives

1. Understand the concept of a sampling distribution.

2. Describe the center, spread, and shape of the sampling distribution of a sample mean.

3. Distinguish between the standard deviation of a population and the standard deviation of a sampling distribution.

4. Explain the content and importance of the Central Limit Theorem.

5. Identify and explain the conditions for using normal approximation involving a sample mean.

6. Verify that the conditions for normal approximation are met and carry out normal approximation involving a sample mean or sample sum.

4.2.1   The mean and standard deviation of $\bar{x}$

In this section we consider a data set called `run17`, which represents all 19,961 runners who finished the 2017 Cherry Blossom 10 mile run in Washington, DC.[22] Part of this data set is shown in Figure 4.16, and the variables are described in Figure 4.17.

ID	time	age	gender	state
1	92.25	38.00	M	MD
2	106.35	33.00	M	DC
⋮	⋮	⋮	⋮	⋮
16923	122.87	37.00	F	VA
16924	93.30	27.00	F	DC

Figure 4.16: Four observations from the `run17` data set.

variable	description
`time`	Ten mile run time, in minutes
`age`	Age, in years
`gender`	Gender (`M` for male, `F` for female)
`state`	Home state (or country if not from the US)

Figure 4.17: Variables and their descriptions for the `run17` data set.

[22]www.cherryblossom.org

These data are special because they include the results for the entire population of runners who finished the 2017 Cherry Blossom Run. We took a simple random sample of this population, which is represented in Figure 4.18. A histogram summarizing the time variable in the `run17samp` data set is shown in Figure 4.19.

ID	time	age	gender	state
1983	88.31	59	M	MD
8192	100.67	32	M	VA
⋮	⋮	⋮	⋮	⋮
1287	89.49	26	M	DC

Figure 4.18: Three observations for the `run17samp` data set, which represents a simple random sample of 100 runners from the 2017 Cherry Blossom Run.
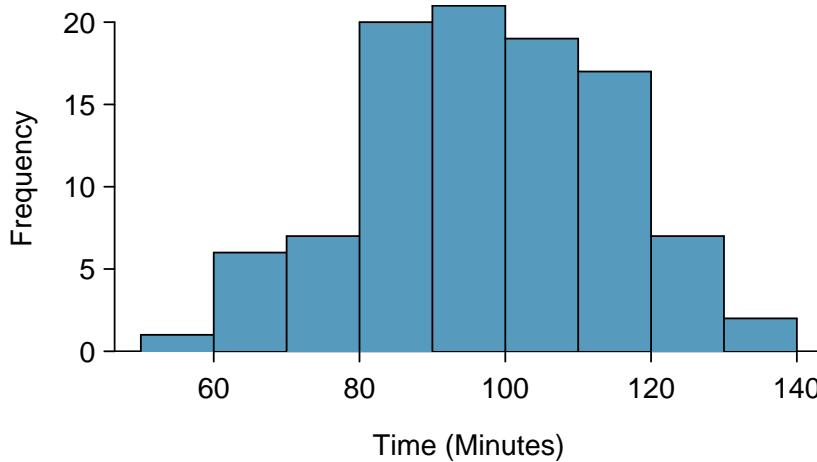
Figure 4.19: Histogram of `time` for a single sample of size 100. The average of the sample is in the mid-90s and the standard deviation of the sample $s \approx 17$ minutes.

From the random sample represented in `run17samp`, we guessed the average time it takes to run 10 miles is 95.61 minutes. Suppose we take another random sample of 100 individuals and take its mean: 95.30 minutes. Suppose we took another (93.43 minutes) and another (94.16 minutes), and so on. If we do this many many times – which we can do only because we have the entire population data set – we can build up a **sampling distribution** for the sample mean when the sample size is 100, shown in Figure 4.20.

SAMPLING DISTRIBUTION

The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a point estimate as being drawn from such a distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

The sampling distribution shown in Figure 4.20 is unimodal and approximately symmetric. It is also centered exactly at the true population mean: $\mu = 94.52$. Intuitively, this makes sense. The sample mean should be an unbiased estimator of the population mean. Because we are considering the distribution of the sample mean, we will use $\mu_{\bar{x}} = 94.52$ to describe the true mean of this distribution.
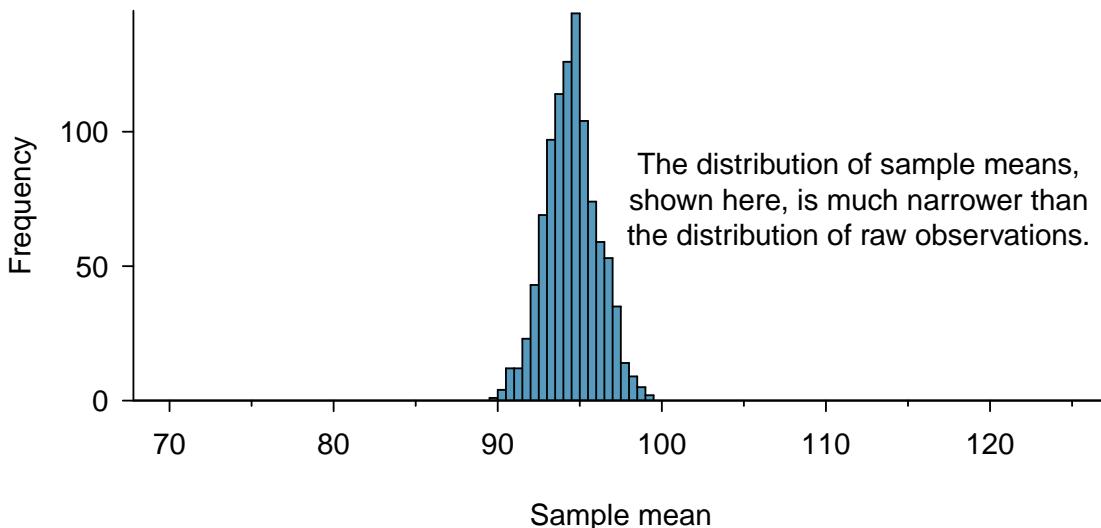
Figure 4.20: A histogram of 1000 sample means for run time, where the samples
are of size $n = 100$. This histogram approximates the true sampling distribution
of the sample mean, with mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$.

We can see that the sample mean has some variability around the population mean, which
can be quantified using the standard deviation of this distribution of sample means. The standard
deviation of the sample mean tells us how far the typical estimate is away from the actual population
mean, 94.52 minutes. It also describes the typical **error** of a single estimate, and is denoted by the
symbol $\sigma_{\bar{x}}$.

STANDARD DEVIATION OF AN ESTIMATE

The standard deviation associated with an estimate describes the typical error or uncertainty
associated with the estimate.

EXAMPLE 4.28

Looking at Figures 4.19 and 4.20, we see that the standard deviation of the sample mean with
$n = 100$ is much smaller than the standard deviation of a single sample. Interpret this statement
and explain why it is true.

The variation from one sample mean to another sample mean is much smaller than the variation from
one individual to another individual. This makes sense because when we average over 100 values,
the large and small values tend to cancel each other out. While many individuals have a time under
90 minutes, it would be unlikely for the *average* of 100 runners to be less than 90 minutes.

GUIDED PRACTICE 4.29

Ⓖ

(a) Would you rather use a small sample or a large sample when estimating a parameter? Why?
(b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard deviation than a point estimate based on a larger sample?[23]

When considering how to calculate the standard deviation of a sample mean, there is one problem: there is no obvious way to estimate this from a single sample. However, statistical theory provides a helpful tool to address this issue.

In the sample of 100 runners, the standard deviation of the sample mean is equal to one-tenth of the population standard deviation: $15.93/10 = 1.59$. In other words, the standard deviation of the sample mean based on 100 observations is equal to

$$SD_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{15.93}{\sqrt{100}} = 1.59$$

where σ_x is the standard deviation of the individual observations. This is no coincidence. We can show mathematically that this equation is correct when the observations are independent using the probability tools of Section 3.5.

---

[23](a) Consider two random samples: one of size 10 and one of size 1000. Individual observations in the small sample are highly influential on the estimate while in larger samples these individual observations would more often average each other out. The larger sample would tend to provide a more accurate estimate. (b) If we think an estimate is better, we probably mean it typically has less error. Based on (a), our intuition suggests that a larger sample size corresponds to a smaller standard deviation.

COMPUTING SD FOR THE SAMPLE MEAN

Given n independent observations from a population with standard deviation σ, the standard deviation of the sample mean is equal to

$$SD_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.30}$$

A reliable method to ensure sample observations are independent is to conduct a simple random sample consisting of less than 10% of the population.

GUIDED PRACTICE 4.31

The average of the runners' ages is 35.05 years with a standard deviation of $\sigma = 8.97$. A simple random sample of 100 runners is taken. (a) What is the standard deviation of the sample mean? (b) Would you be surprised to get a sample of size 100 with an average of 36 years?[24]

GUIDED PRACTICE 4.32

(a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation of the individual observations is 10, what is our estimate of the standard deviation of the mean when the sample size is 100? What about when it is 400? (c) Explain how your answer to (b) mathematically justifies your intuition in part (a).[25]

---

[24](a) Use Equation (4.30) with the population standard deviation to compute the standard deviation of the sample mean: $SD_{\bar{y}} = 8.97/\sqrt{100} = 0.90$ years. (b) It would not be surprising. 36 years is about 1 standard deviation from the true mean of 35.05. Based on the 68, 95 rule, we would get a sample mean at least this far away from the true mean approximately $100\% - 68\% = 32\%$ of the time.

[25](a) Extra observations are usually helpful in understanding the population, so a point estimate with 400 observations seems more trustworthy. (b) The standard deviation of the mean when the sample size is 100 is given by $SD_{100} = 10/\sqrt{100} = 1$. For 400: $SD_{400} = 10/\sqrt{400} = 0.5$. The larger sample has a smaller standard deviation of the mean. (c) The standard deviation of the mean of the sample with 400 observations is lower than that of the sample with 100 observations. The standard deviation of $\bar{x}$ describes the typical error, and since it is lower for the larger sample, this mathematically shows the estimate from the larger sample tends to be better – though it does not guarantee that every large sample will provide a better estimate than a particular small sample.

4.2.2 Examining the Central Limit Theorem

In Figure 4.20, the sampling distribution of the sample mean looks approximately normally distributed. Will the sampling distribution of a mean always be nearly normal? To address this question, we will investigate three cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *normal* distribution. These distributions are shown in the top panels of Figure 4.21. The uniform distribution is symmetric, and the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers).

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the center and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

GUIDED PRACTICE 4.33

Examine the distributions in each row of Figure 4.21. What do you notice about the sampling distribution of the mean as the sample size, n, becomes larger?[26]

EXAMPLE 4.34

In general, would normal approximation for a sample mean be appropriate when the sample size is at least 30?

Yes, the sampling distributions when $n = 30$ all look very much like the normal distribution.

However, the more non-normal a population distribution, the larger a sample size is necessary for the sampling distribution to look nearly normal.

DETERMINING IF THE SAMPLE MEAN IS NORMALLY DISTRIBUTED

If the population is normal, the sampling distribution of $\bar{x}$ will be normal for any sample size.

The less normal the population, the larger n needs to be for the sampling distribution of $\bar{x}$ to be nearly normal. However, a good rule of thumb is that for almost all populations, the sampling distribution of $\bar{x}$ will be approximately normal if $n \geq 30$.

This brings us to the **Central Limit Theorem**, the most fundamental theorem in Statistics.

CENTRAL LIMIT THEOREM

When taking a random sample of independent observations from a population with a fixed mean and standard deviation, the distribution of $\bar{x}$ approaches the normal distribution as n increases.

---

[26]The normal approximation becomes better as larger samples are used. However, in the case when the population is normally distributed, the normal distribution of the sample mean is normal for all sample sizes.
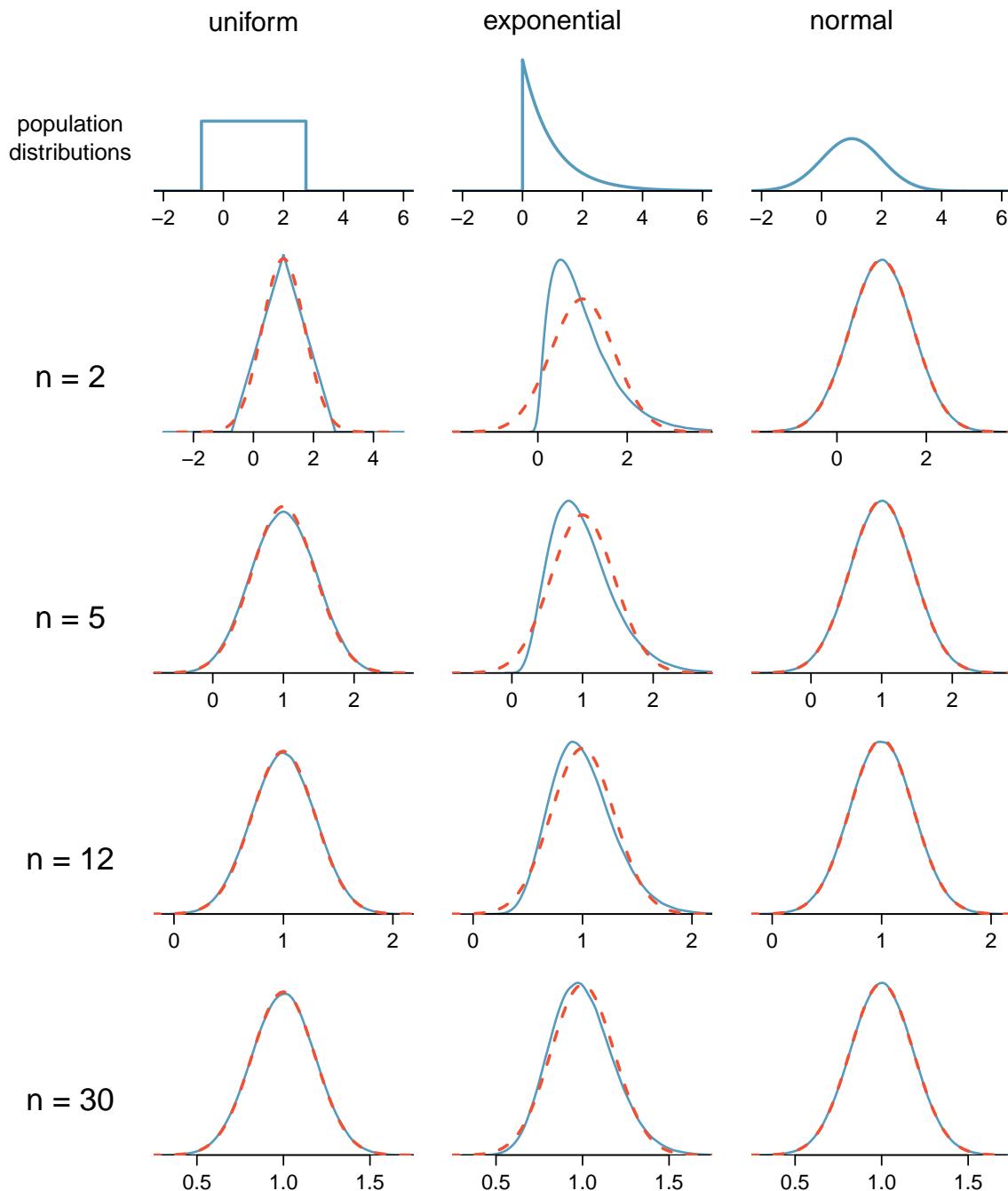
Figure 4.21: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

EXAMPLE 4.35

Sometimes we do not know what the population distribution looks like. We have to infer it based on the distribution of a single sample. Figure 4.22 shows a histogram of 20 observations. These represent winnings and losses from 20 consecutive days of a professional poker player. Based on this sample data, can the normal approximation be applied to the distribution of the sample mean?

————————

We should consider each of the required conditions.

(1) These are referred to as **time series data**, because the data arrived in a particular sequence. If the player wins on one day, it may influence how she plays the next. To make the assumption of independence we should perform careful checks on such data.

(2) The sample size is 20, which is smaller than 30.

(3) There are two outliers in the data, both quite extreme, which suggests the population may not be normal and instead may be very strongly skewed or have distant outliers. Outliers can play an important role and affect the distribution of the sample mean and the estimate of the standard deviation of the sample mean.

Since we should be skeptical of the independence of observations and the extreme upper outliers pose a challenge, we should not use the normal model for the sample mean of these 20 observations. If we can obtain a much larger sample, then the concerns about skew and outliers would no longer apply.
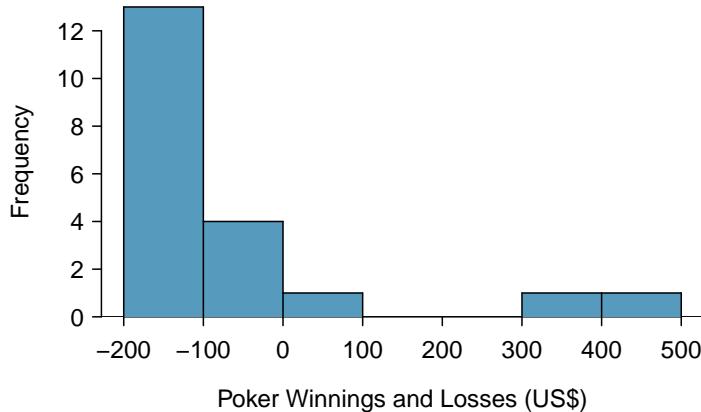
Figure 4.22: Sample distribution of poker winnings. These data include two very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

EXAMINE DATA STRUCTURE WHEN CONSIDERING INDEPENDENCE

Some data sets are collected in such a way that they have a natural underlying structure between observations, e.g. when observations occur consecutively. Be especially cautious about independence assumptions regarding such data sets.

WATCH OUT FOR STRONG SKEW AND OUTLIERS

Strong skew in the population is often identified by the presence of clear outliers in the data. If a data set has prominent outliers, then a larger sample size will be needed for the sampling distribution of $\bar{x}$ to be normal. There are no simple guidelines for what sample size is big enough for each situation. However, we can use the rule of thumb that, in general, an n of at least 30 is sufficient for most cases.

4.2.3   Normal approximation for the sampling distribution of $\bar{x}$

At the beginning of this chapter, we used normal approximation for populations or for data that had an approximately normal distribution. When appropriate conditions are met, we can also use the normal approximation to estimate probabilities about a sample average. We must remember to verify that the conditions are met and use the mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$ for the sampling distribution of the sample average.

THREE IMPORTANT FACTS ABOUT THE DISTRIBUTION OF A SAMPLE MEAN $\bar{x}$

Consider taking a simple random sample from a large population.

1. The mean of a sample mean is denoted by $\mu_{\bar{x}}$, and it is equal to μ.

2. The SD of a sample mean is denoted by $\sigma_{\bar{x}}$, and it is equal to $\frac{\sigma}{\sqrt{n}}$.

3. When the population is normal or when $n \geq 30$, the sample mean closely follows a normal distribution.

EXAMPLE 4.36

In the 2012 Cherry Blossom 10 mile run, the average time for all of the runners is 94.52 minutes with a standard deviation of 8.97 minutes. The distribution of run times is approximately normal. Find the probabiliy that a randomly selected runner completes the run in less than 90 minutes.

Because the distribution of run times is approximately normal, we can use normal approximation.

$$Z = \frac{x - \mu}{\sigma} = \frac{90 - 94.52}{8.97} = -0.504$$
$$P(Z < -0.504) = 0.3072$$

There is a 30.72% probability that a randomly selected runner will complete the run in less than 90 minutes.

EXAMPLE 4.37

Find the probabiliy that the average of 20 runners is less than 90 minutes.

Here, $n = 20 < 30$, but the distribution of the population, that is, the distribution of run times is stated to be approximately normal. Because of this, the sampling distribution will be normal for any sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8.97}{\sqrt{20}} = 2.01$$
$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{90 - 94.52}{2.01} = -2.25$$
$$P(Z < -2.25) = 0.0123$$

There is a 1.23% probability that the average run time of 20 randomly selected runners will be less than 90 minutes.

EXAMPLE 4.38

The average of all the runners' ages is 35.05 years with a standard deviation of $\sigma = 8.97$. The distribution of age is somewhat skewed. What is the probability that a randomly selected runner is older than 37 years?

Because the distribution of age is skewed and is not normal, we cannot use normal approximation for this problem. In order to answer this question, we would need to look at all of the data.

GUIDED PRACTICE 4.39

What is the probability that the average of 50 randomly selected runners is greater than 37 years?[27]

REMEMBER TO DIVIDE BY $\sqrt{n}$

When finding the probability that an *average* or mean is greater or less than a particular value, remember to divide the standard deviation of the population by $\sqrt{n}$ to calculate the correct SD.

---

[27]Because $n = 50 \geq 30$, the sampling distribution of the mean is approximately normal, so we can use normal approximation for this problem. The mean is given as 35.05 years.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8.97}{\sqrt{50}} = 1.27 \qquad z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{37 - 35.05}{1.27} = 1.535 \qquad P(Z > 1.535) = 0.062$$

There is a 6.2% chance that the average age of 50 runners will be greater than 37.
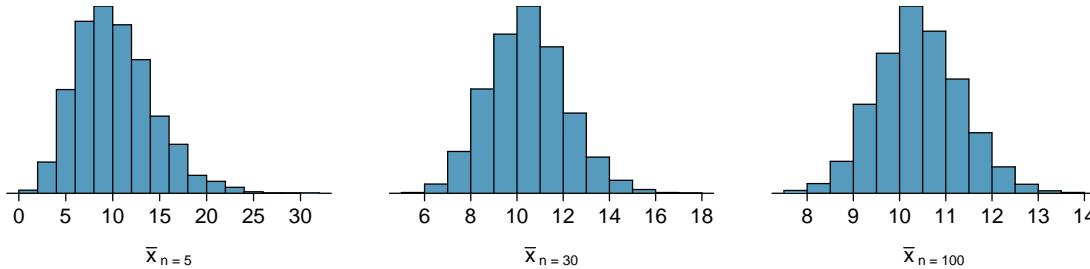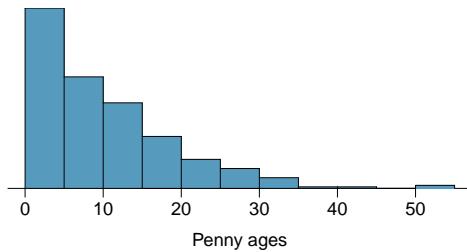
Section summary

- The symbol $\bar{x}$ denotes the sample average. $\bar{x}$ for any particular sample is a number. However, $\bar{x}$ can vary from sample to sample. The distribution of all possible values of $\bar{x}$ for repeated samples of a fixed size from a certain population is called the **sampling distribution** of $\bar{x}$.

- The standard deviation of $\bar{x}$ describes the typical error or distance of the sample mean from the population mean. It also tells us how much the sample mean is likely to vary from one random sample to another.

- The standard deviation of $\bar{x}$ will be *smaller* than the standard deviation of the population by a factor of $\sqrt{n}$. The larger the sample, the better the estimate tends to be.

- Consider taking a simple random sample from a population with a fixed mean and standard deviation. The **Central Limit Theorem** ensures that regardless of the shape of the original population, as the sample size increases, the distribution of the sample average $\bar{x}$ becomes more normal.

- Three important facts about the sampling distribution of the sample average $\bar{x}$:
  - The mean of a sample mean is denoted by $\mu_{\bar{x}}$, and it is equal to μ. (*center*)
  - The SD of a sample mean is denoted by $\sigma_{\bar{x}}$, and it is equal to $\frac{\sigma}{\sqrt{n}}$. (*spread*)
  - When the population is normal or when $n \geq 30$, the sample mean closely follows a normal distribution. (*shape*)

- These facts are used when solving the following two types of **normal approximation** problems involving a *sample mean* or a *sample sum*.

  A: *Find the probability that a sample average will be greater/less than a certain value.*
  1. Verify that the population is approximately normal or that $n \geq 30$.
  2. Calculate the Z-score. Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to standardize the sample average.
  3. Find the appropriate area under the normal curve.

  B: *Find the probability that a sample sum/total will be greater/less than a certain value.*
  1. Convert the sample sum into a sample average, using $\bar{x} = \frac{sum}{n}$.
  2. Do steps 1-3 from Part A above.

Exercises

4.17 Ages of pennies, Part I. The histogram below shows the distribution of ages of pennies at a bank.

(a) Describe the distribution.

(b) Sampling distributions for means from simple random samples of 5, 30, and 100 pennies is shown in the histograms below. Describe the shapes of these distributions and comment on whether they look like what you would expect to see based on the Central Limit Theorem.

Penny ages

$\overline{x}_{n=5}$ $\qquad$ $\overline{x}_{n=30}$ $\qquad$ $\overline{x}_{n=100}$

4.18 Ages of pennies, Part II. The mean age of the pennies from Exercise 4.17 is 10.44 years with a standard deviation of 9.2 years. Using the Central Limit Theorem, calculate the means and standard deviations of the distribution of the mean from random samples of size 5, 30, and 100. Comment on whether the sampling distributions shown in Exercise 4.17 agree with the values you compute.
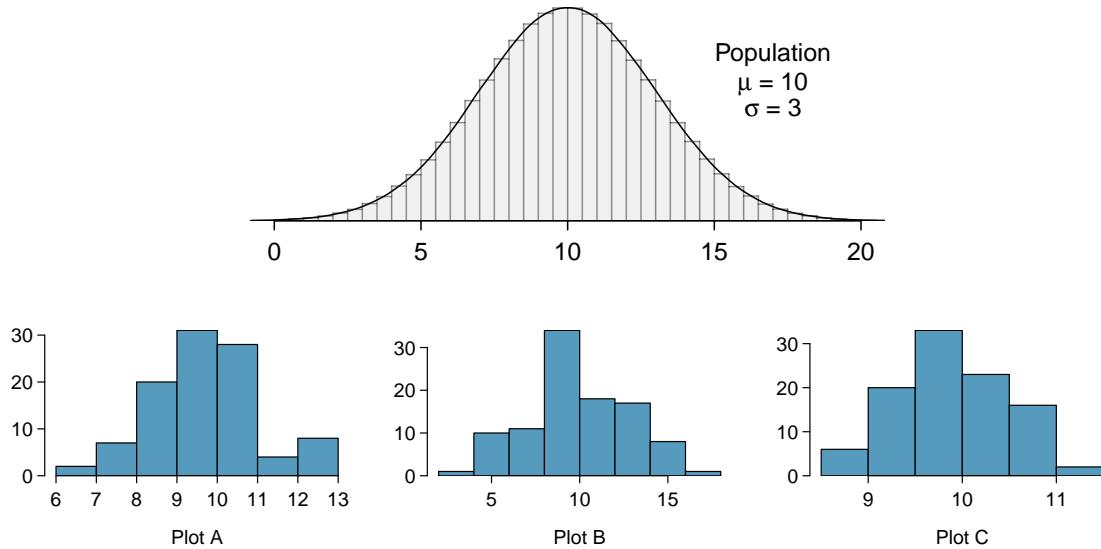
4.19 Housing prices, Part I. ▶️ A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

(a) Is the distribution of housing prices in Topanga symmetric, right skewed, or left skewed? *Hint:* Sketch the distribution.

(b) Would you expect most houses in Topanga to cost more or less than $1.3 million?

(c) Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

(d) What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

(e) How would doubling the sample size affect the standard deviation of the mean?
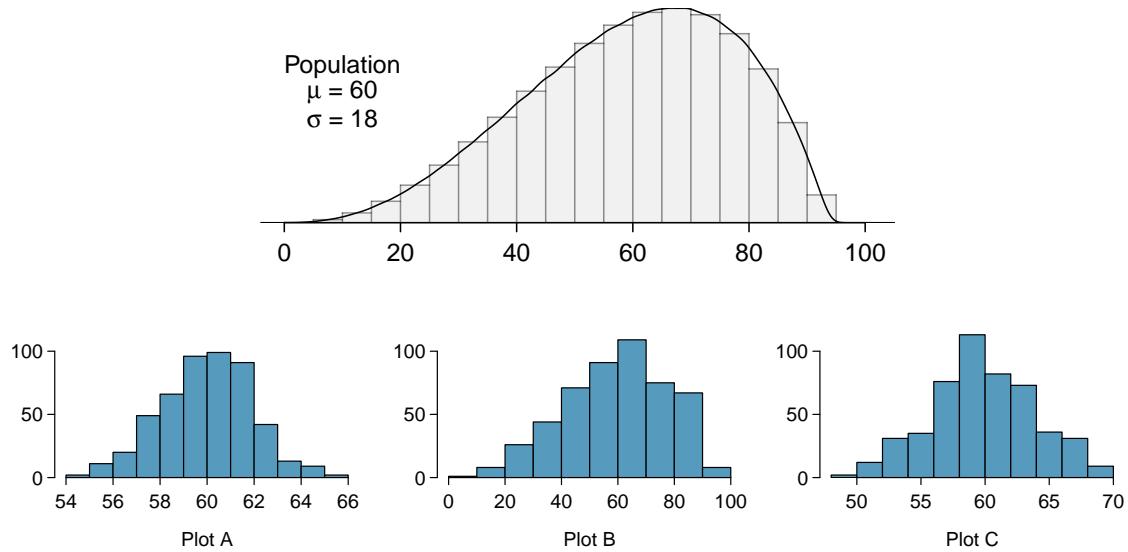
4.20 Stats final scores. Each year about 1500 students take the introductory statistics course at a large university. This year scores on the final exam are distributed with a median of 74 points, a mean of 70 points, and a standard deviation of 10 points. There are no students who scored above 100 (the maximum score attainable on the final) but a few students scored below 20 points.

(a) Is the distribution of scores on this final exam symmetric, right skewed, or left skewed?

(b) Would you expect most students to have scored above or below 70 points?

(c) Can we calculate the probability that a randomly chosen student scored above 75 using the normal distribution?

(d) What is the probability that the average score for a random sample of 40 students is above 75?

(e) How would cutting the sample size in half affect the standard deviation of the mean?

4.21   Identify distributions, Part I. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 10 and the standard deviation is 3. Also shown below is a distribution of (1) a single random sample of 100 values from this population, (2) a distribution of 100 sample means from random samples with size 5, and (3) a distribution of 100 sample means from random samples with size 25. Determine which plot (A, B, or C) is which and explain your reasoning.

Plot A   Plot B   Plot C

4.22   Identify distributions, Part II. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 60 and the standard deviation is 18. Also shown below is a distribution of (1) a single random sample of 500 values from this population, (2) a distribution of 500 sample means from random samples of each size 18, and (3) a distribution of 500 sample means from random samples of each size 81. Determine which plot (A, B, or C) is which and explain your reasoning.

Plot A   Plot B   Plot C
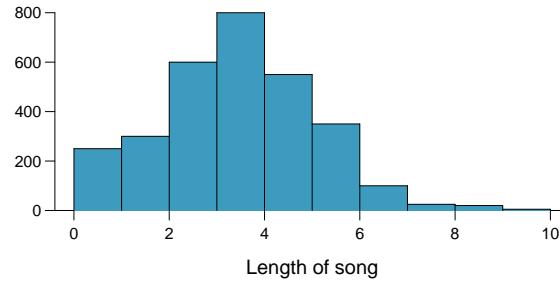
4.23   Weights of pennies. The distribution of weights of United States pennies is approximately normal with a mean of 2.5 grams and a standard deviation of 0.03 grams.

(a) What is the probability that a randomly chosen penny weighs less than 2.4 grams?
(b) Describe the sampling distribution of the mean weight of 10 randomly chosen pennies.
(c) What is the probability that the mean weight of 10 pennies is less than 2.4 grams?
(d) Sketch the two distributions (population and sampling) on the same scale.
(e) Could you estimate the probabilities from (a) and (c) if the weights of pennies had a skewed distribution?

4.24 CFLs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

(b) Describe the distribution of the mean lifespan of 15 light bulbs.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

4.25 Songs on an iPod. Suppose an iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes.

(a) Estimate the probability that a randomly selected song lasts more than 5 minutes.

(b) You are about to go for an hour run and you make a random playlist of 15 songs. What is the probability that your playlist lasts for the entire duration of your run? *Hint:* If you want the playlist to last 60 minutes, what should be the minimum average length of a song?

(c) You are about to take a trip to visit your parents and the drive is 6 hours. You make a random playlist of 100 songs. What is the probability that your playlist lasts the entire drive?

4.26 Spray paint, Part II. As described in Exercise 4.14, the area that can be painted using a single can of spray paint is slightly variable and follows a nearly normal distribution with a mean of 25 square feet and a standard deviation of 3 square feet.

(a) What is the probability that the area covered by a can of spray paint is more than 27 square feet?

(b) Suppose you want to spray paint an area of 540 square feet using 20 cans of spray paint. On average, how many square feet must each can be able to cover to spray paint all 540 square feet?

(c) What is the probability that you can cover a 540 square feet area using 20 cans of spray paint?

(d) If the area covered by a can of spray paint had a slightly skewed distribution, could you still calculate the probabilities in parts (a) and (c) using the normal distribution?

4.27 Wireless routers. John is shopping for wireless routers and is overwhelmed by the number of available options. In order to get a feel for the average price, he takes a random sample of 75 routers and finds that the average price for this sample is $75 and the standard deviation is $25.

(a) Based on this information, how much variability should he expect to see in the mean prices of repeated samples, each containing 75 randomly selected wireless routers?

(b) A consumer website claims that the average price of routers is $80. Is a true average of $80 consistent with John's sample?

4.28 Betting on dinner, Part II. Exercise 4.16 introduces a promotion at a restaurant where prices of menu items are determined randomly following some underlying distribution. We are told that the price of basket of fries is drawn from a normal distribution with mean 6 and standard deviation of 2. You want to get 5 baskets of fries but you only have $28 in your pocket. What is the probability that you would have enough money to pay for all five baskets of fries?

4.3   Geometric distribution

How many times should we expect to roll a die until we get a `1`? How many people should we expect to see at a hospital until we get someone with blood type O+? These questions can be answered using the geometric distribution. We will see that unlike with the distribution of a sample mean, the shape of the geometric distribution is never normal.

Learning objectives

1. Determine if a scenario is geometric.

2. Calculate the probabilities of the possible values of a geometric random variable.

3. Find and interpret the mean (expected value) of a geometric distribution.

4. Understand the shape of the geometric distribution.

4.3.1   Bernoulli distribution

We begin by revisiting a scenario encountered when studying the binomial formula (section 3.3), and we formalize the notion of a yes/no variable.

Many health insurance plans in the United States have a deductible, where the insured individual is responsible for costs up to the deductible, and then the costs above the deductible are shared between the individual and insurance company for the remainder of the year.

Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year. Each of these people can be thought of as a **trial**. We label a person a **success** if her healthcare costs do not exceed the deductible. We label a person a **failure** if she does exceed her deductible in the year. Because 70% of the individuals will not exceed their deductible, we denote the **probability of a success** as $p = 0.7$. The probability of a failure is sometimes denoted with $q = 1 - p$, which would be 0.3 in for the insurance example.

When an individual trial only has two possible outcomes, often labeled as `success` or `failure`, it is called a **Bernoulli random variable**. We chose to label a person who does not exceed her deductible as a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as 1 for a success and 0 for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

$$1\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 0$$

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\#\text{ of successes}}{\#\text{ of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because 0 and 1 are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable.[28]

BERNOULLI RANDOM VARIABLE

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad\qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels 1 and 0 for successes and failures, respectively.

4.3.2 Geometric distribution

The **geometric distribution** is used to describe how many trials it takes to observe a success. Let's first look at an example.

EXAMPLE 4.40

Suppose we are working at the insurance company and need to find a case where the person did not exceed her (or his) deductible as a case study. If the probability a person will not exceed her deductible is 0.7 and we are drawing people at random, what are the chances that the first person will not have exceeded her deductible, i.e. be a success? The second person? The third? What about the probability that we pull $x - 1$ cases before we find the first success, i.e. the first success is the xth person? (If the first success is the fifth person, then we say $x = 5$.)

The probability of stopping after the first person is just the chance the first person will not exceed her (or his) deductible: 0.7. The probability the second person is the first to exceed her deductible is

$$P(\text{second person is the first to exceed deductible})$$
$$= P(\text{the first won't, the second will}) = (0.3)(0.7) = 0.21$$

Likewise, the probability it will be the third case is $(0.3)(0.3)(0.7) = 0.063$.

If the first success is on the xth person, then there are $x - 1$ failures and finally 1 success, which corresponds to the probability $(0.3)^{x-1}(0.7)$. This is the same as $(1 - 0.7)^{x-1}(0.7)$.

---

[28]If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\mu = E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1)$$
$$= 0 \cdot (1 - p) + 1 \cdot p = 0 + p = p$$

Similarly, the variance of X can be computed:

$$\sigma^2 = (0 - p)^2 \cdot P(X = 0) + (1 - p)^2 \cdot P(X = 1)$$
$$= p^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p(1 - p)$$

The standard deviation is $\sigma = \sqrt{p(1 - p)}$.

Example 4.40 illustrates what the **geometric distribution**, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 4.40 is shown in Figure 4.23. In general, the probabilities for a geometric distribution decrease **exponentially** fast.
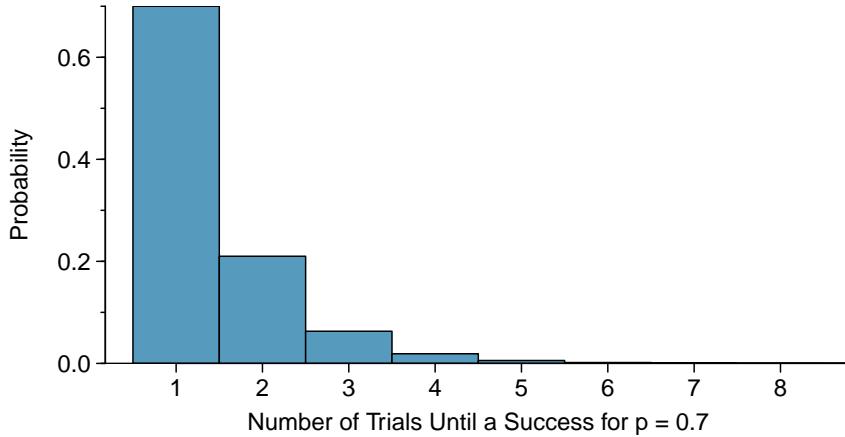
Figure 4.23: The geometric distribution when the probability of success is $p = 0.7$.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation or variance of this distribution, we present general formulas for each.

GEOMETRIC DISTRIBUTION

Let X have a geometric distribution with one parameter p, where p is the probability of a success in one trial. Then the probability of finding the first success in the xth trial is given by

$$P(X = x) = (1 - p)^{x-1} p$$

where $x = 1, 2, 3, \ldots$

The mean (i.e. expected value) and standard deviation of this wait time are given by

$$\mu_X = \frac{1}{p} \qquad \sigma_X = \frac{\sqrt{1 - p}}{p}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

It takes, on average, $1/p$ trials to get a success under the geometric distribution. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

GUIDED PRACTICE 4.41

The probability that a particular case would not exceed their deductible is said to be 0.7. If we were to examine cases until we found one that where the person did not exceed her deductible, how many cases should we expect to check?[29]

EXAMPLE 4.42

What is the chance that we would find the first success within the first 3 cases?
––––––––––––

This is the chance the first ($X = 1$), second ($X = 2$), or third ($X = 3$) case is the first success, which are three disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

$$P(X = 1, 2, \text{ or } 3)$$
$$= P(X = 1) + P(X = 2) + P(X = 3)$$
$$= (0.3)^{1-1}(0.7) + (0.3)^{2-1}(0.7) + (0.3)^{3-1}(0.7)$$
$$= 0.973$$

There is a probability of 0.973 that we would find a successful case within 3 cases.

GUIDED PRACTICE 4.43

Determine a more clever way to solve Example 4.42. Show that you get the same result.[30]

EXAMPLE 4.44

Suppose a car insurer has determined that 88% of its drivers will not exceed their deductible in a given year. If someone at the company were to randomly draw driver files until they found one that had not exceeded their deductible, what is the expected number of drivers the insurance employee must check? What is the standard deviation of the number of driver files that must be drawn?
––––––––––––

In this example, a success is again when someone will not exceed the insurance deductible, which has probability $p = 0.88$. The expected number of people to be checked is $1/p = 1/0.88 = 1.14$ and the standard deviation is $\frac{\sqrt{1-p}}{p} = \frac{\sqrt{1-0.88}}{0.88} = 0.39$.

GUIDED PRACTICE 4.45

Using the results from Example 4.44, $\mu_x = 1.14$ and $\sigma_x = 0.39$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[31]

The independence assumption is crucial to the geometric distribution's accurate description of a scenario. Mathematically, we can see that to construct the probability of the success on the xth trial, we had to use the General Multiplication Rule for independent processes. It is no simple task to generalize the geometric model for dependent trials.

––––––––––––

[29]We would expect to see about $1/0.7 \approx 1.43$ individuals to find the first success.
[30]First find the probability of the complement: $P(\text{no success in first 3 trials}) = 0.3^3 = 0.027$. Next, compute one minus this probability: $1 - P(\text{no success in 3 trials}) = 1 - 0.027 = 0.973$.
[31]No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.

Section summary

- It is useful to model yes/no, success/failure with the values 1 and 0, respectively. We call the **probability of success** p and the **probability of failure** $1 - p$.

- When the trials are **independent** and the value of p is constant, the probability of finding **the first success on the xth trial** is given by $(1 - p)^{x-1}p$. We can see the reasoning behind this formula as follows: for the first success to happen on the xth trial, it has to *not* happen the first $x - 1$ trials (with probability $1 - p$), and then happen on the xth trial (with probability p).

- When we consider the *entire distribution* of possible values for the how long *until* the first success, we get a discrete probability distribution known as the geometric distribution. The **geometric distribution** describes the waiting time *until* the first success, when the trials are independent and the probability of success, p, is constant. If X has a geometric distribution with parameter p, then $P(X = x) = (1 - p)^{x-1}p$, where $x = 1, 2, 3 \ldots$.

- The geometric distribution is always *right skewed* and, in fact, has no maximum value. The probabilities, though, decrease exponentially fast.

- Even though the geometric distribution has an infinite number of values, it has a well-defined **mean**: $\mu_X = \frac{1}{p}$ and **standard deviation**: $\sigma_X = \frac{\sqrt{1-p}}{p}$. If the probability of success is $\frac{1}{10}$, then *on average* it takes 10 trials until we see the first success.

- Note that when the trials are not independent, we can modify the geometric formula to find the probability that the first success happens on the xth trial. Instead of simply raising $(1 - p)$ to the $x - 1$, multiply the appropriate *conditional* probabilities.

Exercises

4.29   Is it Bernoulli?   Determine if each trial can be considered an independent Bernoulli trial for the following situations.

(a) Cards dealt in a hand of poker.

(b) Outcome of each roll of a die.

4.30   With and without replacement.   In the following situations assume that half of the specified population is male and the other half is female.

(a) Suppose you're sampling from a room with 10 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

(b) Now suppose you're sampling from a stadium with 10,000 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

(c) We often treat individuals who are sampled from a large population as independent. Using your findings from parts (a) and (b), explain whether or not this assumption is reasonable.

4.31   Eye color, Part I.   A husband and wife both have brown eyes but carry genes that make it possible for their children to have brown eyes (probability 0.75), blue eyes (0.125), or green eyes (0.125).

(a) What is the probability the first blue-eyed child they have is their third child? Assume that the eye colors of the children are independent of each other.

(b) On average, how many children would such a pair of parents have before having a blue-eyed child? What is the standard deviation of the number of children they would expect to have until the first blue-eyed child?

4.32   Defective rate.   A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10th transistor produced is the first with a defect?

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

4.33   Bernoulli, the mean.   Use the probability rules from Section 3.5 to derive the mean of a Bernoulli random variable, i.e. a random variable X that takes value 1 with probability p and value 0 with probability $1 - p$. That is, compute the expected value of a generic Bernoulli random variable.

4.34   Bernoulli, the standard deviation.   Use the probability rules from Section 3.5 to derive the standard deviation of a Bernoulli random variable, i.e. a random variable X that takes value 1 with probability p and value 0 with probability $1 - p$. That is, compute the square root of the variance of a generic Bernoulli random variable.

4.4   Binomial distribution

What is the probabilty that more than half of a random sample of 40 people would have blood type O+? If the probability of a defective part is 1%, how many defective items would we expect in a random shipment of 200 of those parts? We can model these scenarios and answer these questions using the binomial distribution.

Learning objectives

- Determine if a scenario is binomial.

- Calculate the probabilities of the possible values of a binomial random variable.

- Calculate and interpret the mean (expected value) and standard deviation of the number of successes in n binomial trials.

- Determine whether a binomial distribution can be modeled as approximately normal. If so, use normal approximation to estimate cumulative binomial probabilities.

4.4.1   An example of a binomial distribution

In Guided Practice 3.65, we asked various probability questions regarding the number of people out of 4 with blood type O+. We verified that the scenario was binomial and that each problem could be solved using the binomial formula. Instead of looking at it piecewise, we could describe the entire *distribution* of possible values and their corresponding probabilities. Since there are 4 people, there are several possible outcomes for the number who might have blood type O+: 0, 1, 2, 3, 4. We can make a distribution table with these outcomes. Recall that the probability of a randomly sampled person being blood type O+ is about 0.35.

x_i	$P(x_i)$
0	$\binom{4}{0}(0.35)^0(0.65)^4 = 0.179$
1	$\binom{4}{1}(0.35)^1(0.65)^3 = 0.384$
2	$\binom{4}{2}(0.35)^2(0.65)^2 = 0.311$
3	$\binom{4}{3}(0.35)^3(0.65)^1 = 0.111$
4	$\binom{4}{4}(0.35)^4(0.65)^0 = 0.015$
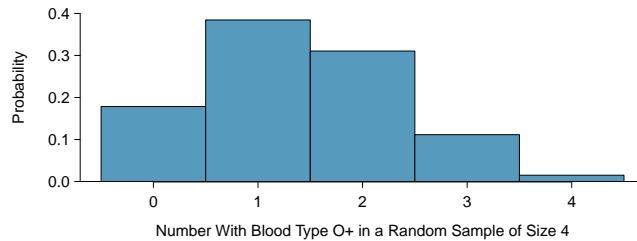
Figure 4.24: Probability distribution for the number with blood type O+ in a random sample of 4 people. This is a binomial distribution. Correcting for rounding error, the probabilities add up to 1, as they must for any probability distribution.

4.4.2   The mean and standard deviation of a binomial distribution

Since this is a probability distribution we could find its mean and standard deviation using the formulas from Chapter 3. Those formulas require a lot of calculations, so it is fortunate there's an easier way to compute the mean and standard deviation for a binomial random variable.

MEAN AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION

For a binomial distribution with parameters n and p, where n is the number of trials and p is the probability of a success, the mean and standard deviation of the number of observed successes are

$$\mu_X = np \qquad\qquad\qquad \sigma_X = \sqrt{np(1-p)}$$

EXAMPLE 4.46

If the probability that a person has blood type O+ is 0.35 and you have 40 randomly selected people, about how many would you expect to have blood type O+? What is the standard deviation of the number of people who would have blood type O+ among the 40 people?

We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas above.

$$\mu_X = np = 40(0.35) = 14$$
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{40(0.35)(0.65)} = 3.0$$

The exact distribution is shown in Figure 4.25.
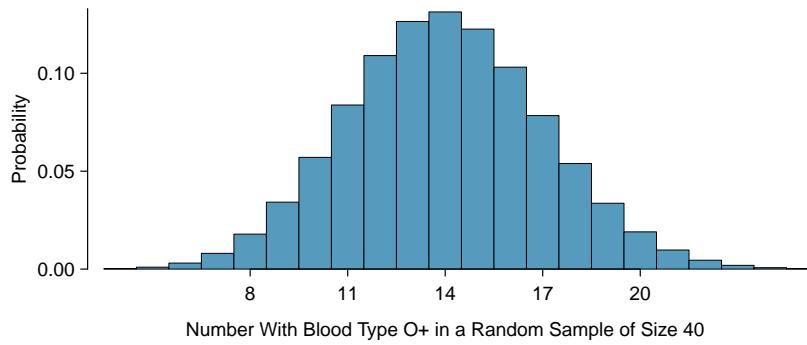
Figure 4.25: Distribution for the number of people with blood type O+ in a random sample of size 40, where $p = 0.35$. The distribution is binomial and is centered on 14 with a standard deviation of 3.

4.4.3   Normal approximation to the binomial distribution

The binomial formula is cumbersome when the sample size (n) is large, particularly when we consider a range of observations.

EXAMPLE 4.47

Find the probability that fewer than 12 out of 40 randomly selected people would have blood type O+, where probability of blood type O+ is 0.35.

This is equivalent to asking, what is the probability of observing $X = 0, 1, 2, ...,$ or 11 with blood type O+ in a sample of size 40 when $p = 0.35$? We previously verified that this scenario is binomial. We can compute each of the 12 probabilities using the binomial formula and add them together to find the answer:

$$P(X = 0 \text{ or } X = 1 \text{ or } \cdots \text{ or } X = 11)$$
$$= P(X = 0) + P(X = 1) + \cdots + P(X = 11)$$
$$= \binom{40}{0}(0.35)^0(0.65)^{40} + \binom{40}{1}(0.35)^1(0.65)^{39} + \cdots + \binom{40}{11}(0.35)^{11}(0.65)^{29}$$
$$= 0.21$$

If the true proportion with blood type O+ in the population is $p = 0.35$, then the probability of observing fewer than 12 in a sample of $n = 40$ is 0.21.

The computations in Example 4.4.3 are tedious and long. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. In some cases we may use the normal distribution to estimate binomial probabilities. While a normal approximation for the distribution in Figure 4.24 when the sample size was $n = 4$ would not be appropriate, it might not be too bad for the distribution in Figure 4.25 where $n = 40$. We might wonder, when is it reasonable to use the normal model to approximate a binomial distribution?

GUIDED PRACTICE 4.48

Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 4.26 shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10, 30, 100, 300$. What happens to the shape of the distributions as the sample size increases? How does the binomial distribution change as n gets larger?[32]

The shape of the binomial distribution depends upon both n and p. Here we introduce a rule of thumb for when normal approximation of a binomial distribution is reasonable. We will use this rule of thumb in many applications going forward.

> **NORMAL APPROXIMATION OF THE BINOMIAL DISTRIBUTION**
>
> The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that $np \geq 10$ and $n(1 - p) \geq 10$. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:
>
> $$\mu = np \qquad\qquad \sigma = \sqrt{np(1 - p)}$$

---

[32]The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in the last hollow histogram.
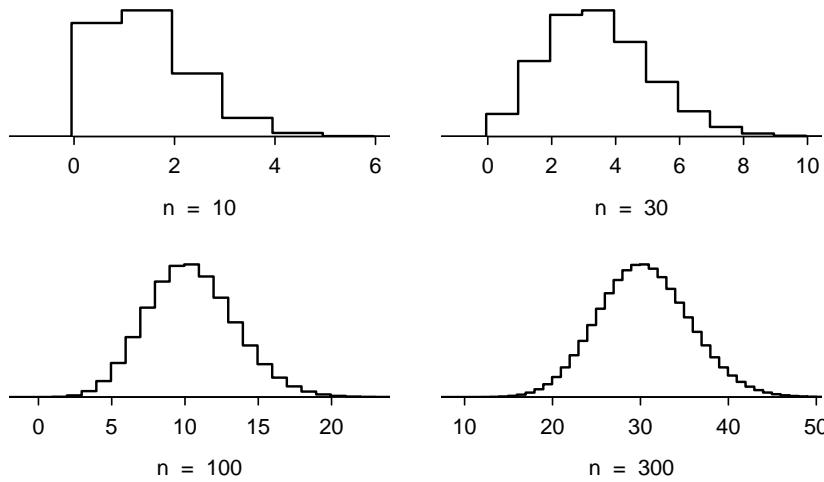
Figure 4.26: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting described in Figure 4.25.

EXAMPLE 4.49

Use the normal approximation to estimate the probability of observing fewer than 12 with blood type O+ in a random sample of 40, if the true proportion with blood type O+ in the population is $p = 0.35$.
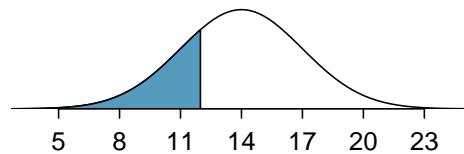
First we verify that np and $n(1 - p)$ are at least 10 so that we can apply the normal approximation to the binomial model:

$$np = 40(0.35) = 14 \geq 10 \qquad\qquad n(1 - p) = 40(0.65) = 26 \geq 10$$

With these conditions checked, we may use the normal distribution to approximate the binomial distribution with the following mean and standard deviation:

$$\mu = np = 40(0.35) = 14$$

$$\sigma = \sqrt{np(1 - p)} = \sqrt{40(0.35)(0.65)} = 3.0$$

We want to find the probability of observing fewer than 12 with blood type O+ using this model. We note that 12 is less than 1 standard deviation below the mean:

Next, we compute the Z-score as $Z = \frac{12-14}{3} = -0.67$ to find the shaded area in the picture: $P(Z < -0.67) = 0.25$. This probability of 0.25 using the normal approximation is reasonably close to the true probability of 0.21 computed using the binomial distribution.

EXAMPLE 4.50

Use the normal approximation to estimate the probability of observing fewer than 120 people with blood type O+ in a random sample of 400, if the true proportion with blood type O+ in the population is $p = 0.35$.

We have previously verified that the binomial model is reasonable for this context. Now we will verify that both np and $n(1-p)$ are at least 10 so we can apply the normal approximation to the binomial model:
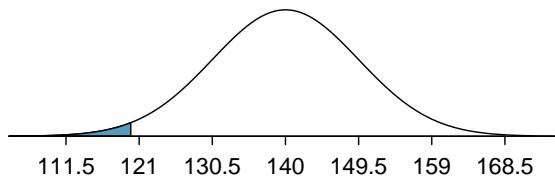
$$np = 400(0.35) = 140 \geq 10 \qquad\qquad n(1-p) = 400(0.35) = 260 \geq 10$$

With these conditions checked, we may use the normal approximation in place of the binomial distribution with the following mean and standard deviation:

$$\mu = np = 400(0.35) = 140$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{400(0.35)(0.65)} = 9.5$$

We want to find the probability of observing fewer than 120 with blood type O+ using this model. We note that 120 is just over 2 standard deviations below the mean:

Next, we compute the Z-score as $Z = \frac{120-140}{9.5} = -2.1$ to find the shaded area in the picture: $P(Z < -2.1) = 0.0179$. This probability of 0.0179 using the normal approximation is very close to the true probability of 0.0196 from the binomial distribution.

GUIDED PRACTICE 4.51

Use normal approximation, if applicable, to estimate the probability of getting greater than 15 sixes in 100 rolls of a fair die.[33]

4.4.4   Normal approximation breaks down on small intervals (special topic)

THE NORMAL APPROXIMATION MAY FAIL ON SMALL INTERVALS

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

We consider again our example where 35% of people are blood type O+. Suppose we want to find the probability that between 129 and 131 people, inclusive, have blood type O+ in a random sample of 400 people. We want to compute the probability of observing 129, 130, or 131 people with blood type O+ when $p = 0.20$ and $n = 400$. With such a large sample, we might be tempted to apply the normal approximation and use the range 129 to 131. However, we would find that the

---

[33]$np = 100(1/6) = 16.7 \geq 10$ and $n(1-p) = 100(5/6) = 83.3 \geq 10$
$\mu = np = 100(1/6) = 16.7; \sigma = \sqrt{np(1-p)} = \sqrt{100(1/6)(5/6)} = 3.7$
$Z = \frac{15-16.7}{3.7} = -0.46.$
$P(Z > -0.46) = 0.677$

binomial solution and the normal approximation notably differ:

<div align="center">

Binomial: 0.0732   Normal: 0.0483

</div>

We can identify the cause of this discrepancy using Figure 4.27, which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval. The binomial distribution is a discrete distribution, and the each bar is centered over an integer value. Looking closely at Figure 4.27, we can see that the bar corresponding to 129 begins at 128.5 and ends at 129.5, the bar corresponding to 131 begins at 130.5 and ends at 131.5, etc.
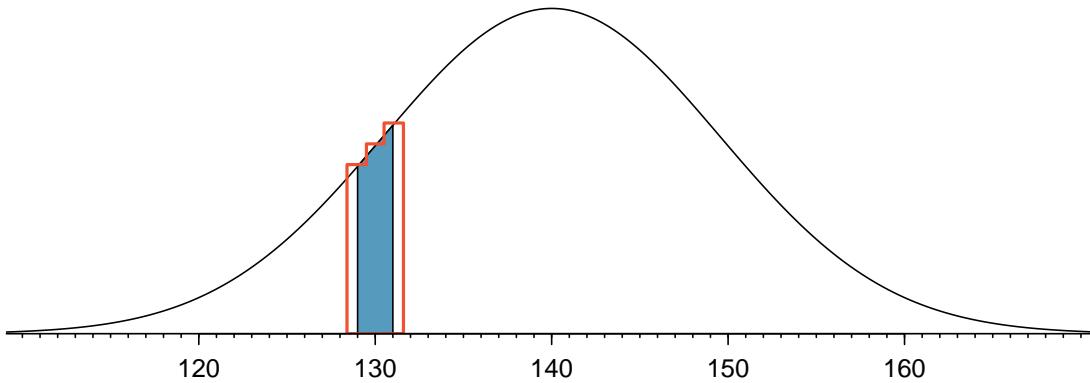
Figure 4.27: A normal curve with the area between 129 and 131 shaded. The outlined area from 128.5 to 131.5 represents the exact binomial probability.

IMPROVING ACCURACY OF THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values for the lower end of a shaded region are reduced by 0.5 and the cutoff value for the upper end are increased by 0.5. This correction is called the continuity correction and accounts for the fact that the binomial distribution is discrete.

EXAMPLE 4.52

Use the method described to find a more accurate estimate for the probability of observing 129, 130, or 131 people with blood type O+ in 400 randomly selected people when $p = 0.35$.

Instead of standardizing 129 and 131, we will standardize 128.5 and 131.5:

$$Z_{left} = \frac{128.5 - 140}{9.5} = -1.263$$
$$Z_{right} = \frac{131.5 - 140}{9.5} = -0.895$$
$$P(-1.263 < Z < -0.895) = 0.0772$$

The probability 0.0772 is much closer to the true value of 0.0732 than the previous estimate of 0.0483 we calculated using normal approximation without the continuity correction.

It is always possible to apply the continuity correction when finding a normal approximation to the binomial distribution. However, when n is very large or when the interval is wide, the benefit of the modification is limited since the added area becomes negligible compared to the overall area being calculated.

Section summary

In the previous chapter, we introduced the binomial formula to find the probability of exactly x successes in n trials for an event that has probability p of success. Instead of looking at this scenario piecewise, we can describe the entire *distribution* of the number of successes and their corresponding probabilities.

- The distribution of the *number of successes* in n independent trials gives rise to a **binomial distribution**. If X has a binomial distribution with parameters n and p, then
  $P(X = x) = \binom{n}{x} p^x (1 - p)^n - x$, where $x = 0, 1, 2, 3 \ldots, n$.

- To write out a binomial probability **distribution table**, list all possible values for x, the number of successes, then use the binomial formula to find the probability of each of those values.

- Because a binomial distribution can be thought of as the *sum* of a bunch of 0s and 1s, the **Central Limit Theorem** applies. As n gets larger, the shape of the binomial distribution becomes more normal.

- We call the rule of thumb for when the binomial distribution can be well modeled with a normal distribution the **success-failure** condition. The success-failure condition is met when there are at least 10 successes and 10 failures, or when $np \geq 10$ and $n(1 - p) \geq 10$.

- If X follows a binomial distribution with parameters n and p, then:
  - The mean is given by $\mu_X = np$.   (*center*)
  - The standard deviation is given by $\sigma_X = \sqrt{np(1 - p)}$.   (*spread*)
  - When $np \geq 10$ and $n(1 - p) \geq 10$, the binomial distribution is approximately normal. (*shape*)

- It is often easier to use **normal approximation to the binomial distribution** rather than evaluate the binomial formula many times. These three properties of the binomial distribution are used when solving the following type of problem.

  Find the probability of getting more than / fewer than x yeses in n trials or in a sample of size n.

  1. Identify n and p. Verify than $np \geq 10$ and $n(1 - p) \geq 10$, which implies that normal approximation is reasonable.
  2. Calculate the Z-score. Use $\mu_X = np$ and $\sigma_X = \sqrt{np(1 - p)}$ to standardize the x value.
  3. Find the appropriate area under the normal curve.

Exercises

4.35   Underage drinking, Part II. ▶️   We learned in Exercise 3.31 that about 70% of 18-20 year olds consumed alcoholic beverages in any given year. We now consider a random sample of fifty 18-20 year olds.

(a) How many people would you expect to have consumed alcoholic beverages? And with what standard deviation?

(b) Would you be surprised if there were 45 or more people who have consumed alcoholic beverages?

(c) What is the probability that 45 or more people in this sample have consumed alcoholic beverages? How does this probability relate to your answer to part (b)?

4.36   Chickenpox, Part II. We learned in Exercise 3.32 that about 90% of American adults had chickenpox before adulthood. We now consider a random sample of 120 American adults.

(a) How many people in this sample would you expect to have had chickenpox in their childhood? And with what standard deviation?

(b) Would you be surprised if there were 105 people who have had chickenpox in their childhood?

(c) What is the probability that 105 or fewer people in this sample have had chickenpox in their childhood? How does this probability relate to your answer to part (b)?

4.37   Game of dreidel. A dreidel is a four-sided spinning top with the Hebrew letters *nun*, *gimel*, *hei*, and *shin*, one on each side. Each side is equally likely to come up in a single spin of the dreidel. Suppose you spin a dreidel three times. Calculate the probability of getting

(a) at least one *nun*?

(b) exactly 2 *nun*s?

(c) exactly 1 *hei*?

(d) at most 2 *gimel*s?

Photo by Staccabees, cropped
(http://flic.kr/p/7gLZTf)
CC BY 2.0 license

4.38   Sickle cell anemia. Sickle cell anemia is a genetic blood disorder where red blood cells lose their flexibility and assume an abnormal, rigid, "sickle" shape, which results in a risk of various complications. If both parents are carriers of the disease, then a child has a 25% chance of having the disease, 50% chance of being a carrier, and 25% chance of neither having the disease nor being a carrier. If two parents who are carriers of the disease have 3 children, what is the probability that

(a) two will have the disease?

(b) none will have the disease?

(c) at least one will neither have the disease nor be a carrier?

(d) the first child with the disease will the be 3rd child?

4.5   Sampling distribution of a sample proportion

Often, instead of the number of successes in n trials, we are interested in the *proportion* of successes in n trials. We can use the sampling distribution of a sample proportion to answer questions such as the following:

- Given a population that is 50% male, what is the probability that a random sample of 200 people would consist of less than 45% males?

- In a particular state, 48% support a controversial measure. When estimating the percent through polling, what is the probability that a random sample of size 200 will mistakenly estimate the percent support to be greater than 50%?

Learning objectives

1. Describe the center, spread, and shape of the sampling distribution of a sample proportion.

2. Recognize the relationship between the distribution of a sample proportion and the corresponding binomial distribution.

3. Identify and explain the conditions for using normal approximation involving a sample proportion. Recognize that the Central Limit Theorem applies in the case of proportions/counts as well as means/sums.

4. Verify that the conditions for normal approximation are met and carry out normal approximation involving a sample proportion or sample count.

4.5.1   The mean and standard deviation of $\hat{p}$

To answer these questions, we investigate the distribution of the sample proportion $\hat{p}$. In the last section we saw that the *number* of people with blood type O+ in a random sample of size 40 follows a binomial distribution with $n = 40$ and $p = 0.35$ that is centered on 14 and has standard deviation 3.0. What does the distribution of the *proportion* of people with blood type O+ in a sample of size 40 look like? To convert from a count to a proportion, we divide the count (i.e. number of yeses) by the sample size, $n = 40$. For example, 6 becomes $8/40 = 0.20$ as a proportion and 11 becomes $11/40 = 0.275$.

We can find the general formula for the mean (expected value) and standard deviation of a sample proportion $\hat{p}$ using our tools that we've learned so far. To get the sample mean for $\hat{p}$, we divide the binomial mean $\mu_{binomial} = np$ by n:

$$\mu_{\hat{p}} = \frac{\mu_{binomial}}{n} = \frac{np}{n} = p$$

As one might expect, the sample proportion $\hat{p}$ is centered on the true proportion p. Likewise, the standard deviation of $\hat{p}$ is equal to the standard deviation of the binomial distribution divided by n:

$$\sigma_{\hat{p}} = \frac{\sigma_{binomial}}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

MEAN AND STANDARD DEVIATION OF A SAMPLE PROPORTION

The mean and standard deviation of the sample proportion describe the center and spread of the distribution of all possible sample proportions $\hat{p}$ from a random sample of size n with true population proportion p.

$$\mu_{\hat{p}} = p \qquad\qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In analyses, we think of the formula for the standard deviation of a sample proportion, $\sigma_{\hat{p}}$, as describing the uncertainty associated with the estimate $\hat{p}$. That is, $\sigma_{\hat{p}}$ can be thought of as a way to quantify the typical error in our sample estimate $\hat{p}$ of the true proportion p. Understanding the variability of statistics such as $\hat{p}$ is a central component in the study of statistics.

Here, $n = 40$ and $p = 0.35$, $\sigma_{\hat{p}} = \sqrt{\frac{0.35(1-0.35)}{40}} = 0.075$. We see in Figure 4.28 that the distribution of number of people in a sample with blood type O+ out of 40 is equivalent to the distribution of proportion of people in a sample of size 40 with blood type O+, but with a change of scale. Instead of counts along the horizontal axis, we have proportions.

(a)

(b)
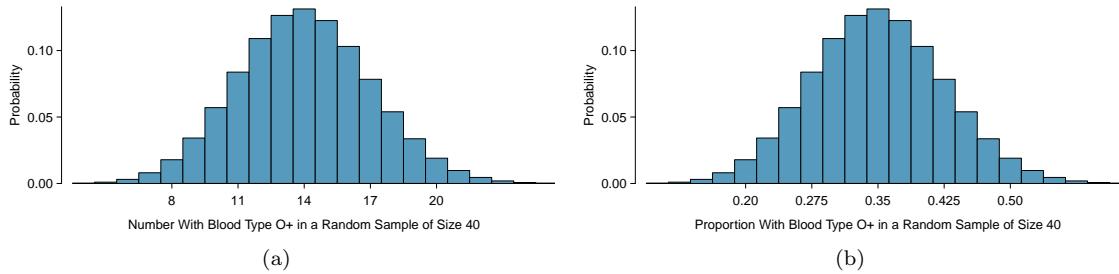
Figure 4.28: Two distributions where $p = 0.35$ and $n = 40$: the binomial distribution for the *number* with blood type O+ and the sampling distribution for the *proportion* with blood type O+.

EXAMPLE 4.53

If the proportion of people in the county with blood type O+ is really 35%, find and interpret the mean and standard deviation of the sample proportion for a random sample of size 400.

The mean of the sample proportion is the population proportion: 0.35. That is, if we took many, many samples and calculated $\hat{p}$, these values would average out to $p = 0.35$.

The standard deviation of $\hat{p}$ is described by the standard deviation for the proportion:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.35(0.65)}{400}} = 0.024$$

The sample proportion will typically be about 0.024 or 2.4% away from the true proportion of $p = 0.35$. We'll become more rigorous about quantifying how close $\hat{p}$ will tend to be to p in Chapter 5.

4.5.2 The Central Limit Theorem revisited

In section 4.2, we saw the Central Limit Theorem, which states that for large enough n, the sample mean $\bar{x}$ is normally distributed.

A natural question is, what does this have to do with sample proportions? In fact, a lot! A sample proportion can be written down as a sample mean. For example, suppose we have 3 successes in 10 trials. If we label each of the 3 success as a 1 and each of the 7 failures as a 0, then the sample

proportion is the same as the sample mean:

$$\hat{p} = \frac{1+0+0+1+1+0+0+0+0+0}{10} = \frac{3}{10} = 0.3$$

That is, the distribution of the sample proportion is governed by the Central Limit Theorem, and the Central Limit Theorem is what ties together much of the statistical theory we will see.

THREE IMPORTANT FACTS ABOUT THE DISTRIBUTION OF A SAMPLE PROPORTION $\hat{p}$

Consider taking a simple random sample from a large population.

1. The mean of a sample proportion is p.

2. The SD of a sample proportion is $\sqrt{\frac{p(1-p)}{n}}$.

3. When $np \geq 10$ and $n(1 - p) \geq 10$, the sample proportion closely follows a normal distribution.

Using these facts, we can now answer the question posed at the beginning of this section.

4.5.3   Normal approximation for the distribution of $\hat{p}$

EXAMPLE 4.54

Find the probability that less than 30% of a random sample of 400 people will be blood type O+ if the population proportion is 35%.

In the previous section we verified that np and $n(1 - p)$ are at least 10. The mean of the sample proportion is 0.35 and the standard deviation for the sample proportion is given by $\sqrt{\frac{0.35(1-0.35)}{400}} = 0.024$. We can find a Z-score and use our calculator to find the probability:

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.30 - 0.35}{0.024} = -2.1$$
$$P(Z < -2.1) = 0.0179$$

We leave it to the reader to construct a figure for this example.

EXAMPLE 4.55

The probability 0.0179 is the same probability we calculated when we found the probability of getting fewer than 120 with blood type O+ out of 400! Why is this?

Notice that $120/400 = 0.30$. Using the binomial distribution to find the probability of fewer than 120 with blood type O+ in the sample is equivalent to using the distribution of $\hat{p}$ to find the probability of a sample proportion less than 0.30.

GUIDED PRACTICE 4.56

Given a population that is 50% male, what is the probability that a sample of size 200 would have greater than 55% males? Remember to verify that conditions for normal approximation are met.[34]

---

[34]First, verify the conditions: $np = 200(0.5) = 100 \geq 10$ and $n(1 - p) = 200(0.5) = 100 \geq 10$, so the normal approximation is reasonable. Next we find the mean and standard deviation of $\hat{p}$:

$$\mu_{\hat{p}} = p = 0.50 \qquad\qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(0.5)}{200}} = 0.0354$$

Section summary

The binomial distribution shows the distribution of the number of successes in n trials. Often, we are interested in the *proportion* of successes rather than the *number* of successes.

- To convert from "number of yeses" to "proportion of yeses" we simply divide the number by n. The sampling distribution of the sample proportion $\hat{p}$ is identical to the binomial distribution with a change of scale, i.e. different mean and different SD, but same shape.

- The same **success-failure condition** for the binomial distribution holds for a sample proportion $\hat{p}$.

- Three important facts about the sampling distribution of the sample proportion $\hat{p}$:

  - The mean of a sample proportion is denoted by $\mu_{\hat{p}}$, and it is equal to p. (*center*)

  - The SD of a sample proportion is denoted by $\sigma_{\hat{p}}$, and it is equal to $\sqrt{\frac{p(1-p)}{n}}$. (*spread*)

  - When $np \geq 10$ and $n(1-p) \geq 10$, the distribution of the sample proportion will be approximately normal. (*shape*)

- We use these properties when solving the following type of **normal approximation** problem involving a sample proportion. *Find the probability of getting more / less than x% yeses in a sample of size n.*

  1. Identify n and p. Verify than $np \geq 10$ and $n(1-p) \geq 10$, which implies that normal approximation is reasonable.

  2. Calculate the Z-score. Use $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to standardize the sample proportion.

  3. Find the appropriate area under the normal curve.
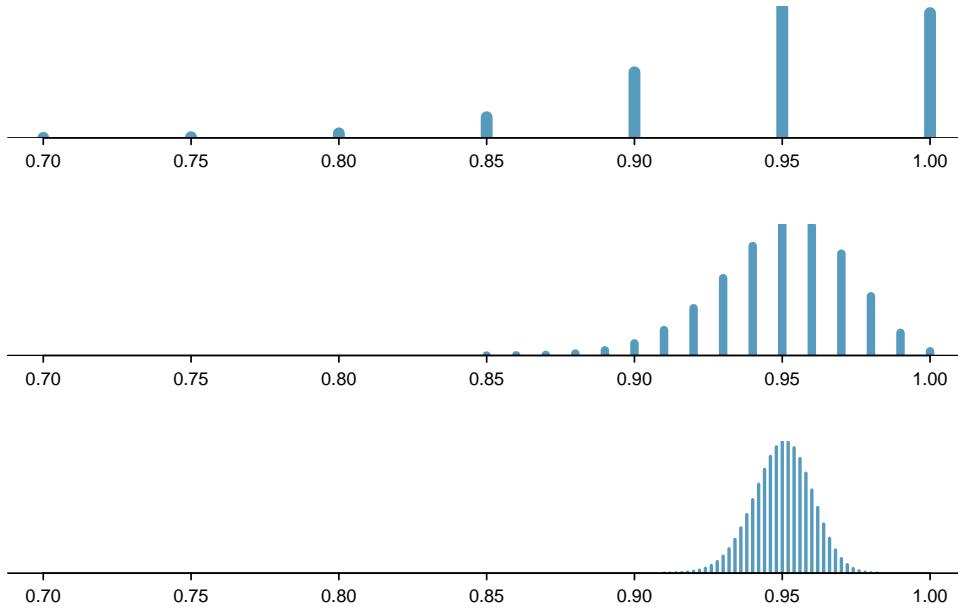
Then we find a Z-score and find the upper tail of the normal distribution:

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.55 - 0.5}{0.0354} = 1.412 \qquad \rightarrow \qquad P(Z > 1.412) = 0.07$$
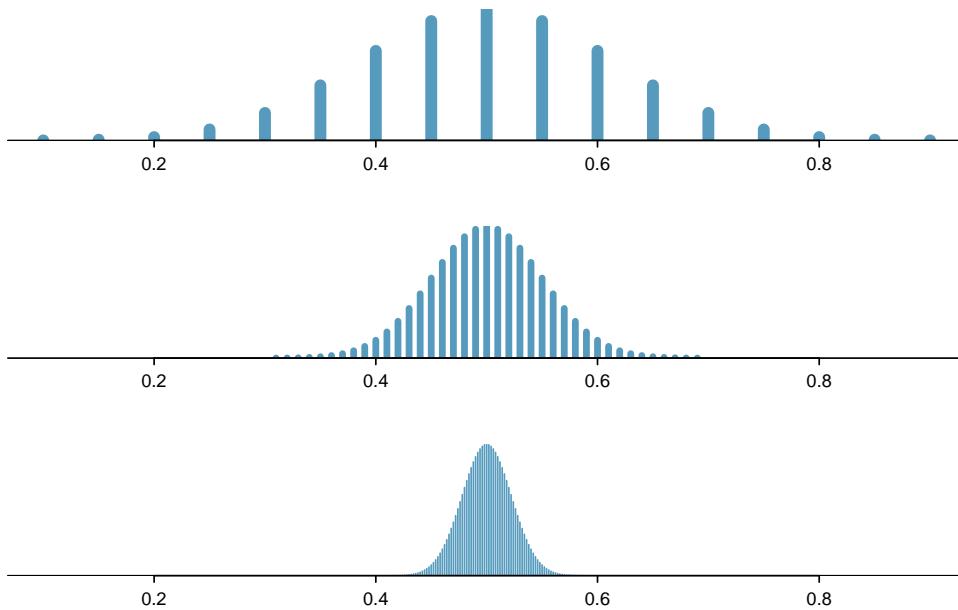
The probability of getting a sample proportion of 55% or greater is about 0.07.

Exercises

4.39   Distribution of $\hat{p}$. Suppose the true population proportion were $p = 0.95$. The figure below shows
what the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$.
(a) What does each point (observation) in each of the samples represent? (b) Describe the distribution of
the sample proportion, $\hat{p}$. How does the distribution of the sample proportion change as n becomes larger?

4.40   Distribution of $\hat{p}$. Suppose the true population proportion were $p = 0.5$. The figure below shows what
the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$.
What does each point (observation) in each of the samples represent? Describe how the distribution of the
sample proportion, $\hat{p}$, changes as n becomes larger.

4.41   Distribution of $\hat{p}$. Suppose the true population proportion were $p = 0.5$ and a researcher takes a simple random sample of size $n = 50$.

(a) Find and interpret the standard deviation of the sample proportion $\hat{p}$.

(b) Calculate the probability that the sample proportion will be larger than 0.55 for a random sample of size 50.

4.42   Distribution of $\hat{p}$. Suppose the true population proportion were $p = 0.6$ and a researcher takes a simple random sample of size $n = 50$.

(a) Find and interpret the standard deviation of the sample proportion $\hat{p}$.

(b) Calculate the probability that the sample proportion will be larger than 0.65 for a random sample of size 50.

4.43   Nearsighted children. It is believed that nearsightedness affects about 8% of all children. We are interested in finding the probability that fewer than 12 out of 200 randomly sampled children will be nearsighted.

(a) Estimate this probability using the normal approximation to the binomial distribution.

(b) Estimate this probability using the distribution of the sample proportion.

(c) How do your answers from parts (a) and (b) compare?

4.44   Social network use. The Pew Research Center estimates that as of January 2014, 89% of 18-29 year olds in the United States use social networking sites.[35] Calculate the probability that at least 95% of 500 randomly sampled 18-29 year olds use social networking sites.

---

[35]Pew Research Center, Washington, D.C. Social Networking Fact Sheet, accessed on May 9, 2015.

Chapter highlights

This chapter began by introducing the normal distribution. A common thread that ran through this chapter is the use of the **normal approximation** in various contexts. The key steps are included for each of the normal approximation scenarios below.

1. Normal approximation for **data**:
   - Verify that population is approximately normal.
   - Use the given mean μ and SD σ to find the Z-score for the given x value.

2. Normal approximation for a **sample mean/sum**:
   Verify that population is approximately normal or that $n \geq 30$.
   Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to find the Z-score for the given/calculated sample mean.

3. Normal approximation for the **number of successes** (binomial):
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$ to find the Z-score for the given number of successes.

4. Normal approximation for a **sample proportion**:
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to find the Z-score for the given sample proportion.

5. Normal approximation for the **sum of two independent random variables**:
   - Verify that each random variable is approximately normal.
   - Use $E(X + Y) = E(X) + E(Y)$ and $SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$ to find the Z-score for the given sum.

Cases 1 and 2 apply to **numerical** variables, while cases 3 and 4 are for **categorical** yes/no variables. Case 5 applies to both numerical and categorical variables.

Note that in the binomial case and in the case of proportions, we never look to see if the *population* is normal. That would not make sense because the "population" is simply a bunch of no/yes, 0/1 values and could not possibly be normal.

The **Central Limit Theorem** is the mathematical rule that ensures that when the sample size is sufficiently large, the sample mean/sum and sample proportion/count will be approximately normal.

Chapter exercises

4.45   Roulette winnings. In the game of roulette, a wheel is spun and you place bets on where it will stop. One popular bet is that it will stop on a red slot; such a bet has an 18/38 chance of winning. If it stops on red, you double the money you bet. If not, you lose the money you bet. Suppose you play 3 times, each time with a $1 bet. Let Y represent the total amount won or lost. Write a probability model for Y.

4.46   Speeding on the I-5, Part I. The distribution of passenger vehicle speeds traveling on the Interstate 5 Freeway (I-5) in California is nearly normal with a mean of 72.6 miles/hour and a standard deviation of 4.78 miles/hour.[36]

(a) What percent of passenger vehicles travel slower than 80 miles/hour?

(b) What percent of passenger vehicles travel between 60 and 80 miles/hour?

(c) How fast do the fastest 5% of passenger vehicles travel?

(d) The speed limit on this stretch of the I-5 is 70 miles/hour. Approximate what percentage of the passenger vehicles travel above the speed limit on this stretch of the I-5.

4.47   University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

4.48   Speeding on the I-5, Part II. Exercise 4.46 states that the distribution of speeds of cars traveling on the Interstate 5 Freeway (I-5) in California is nearly normal with a mean of 72.6 miles/hour and a standard deviation of 4.78 miles/hour. The speed limit on this stretch of the I-5 is 70 miles/hour.

(a) A highway patrol officer is hidden on the side of the freeway. What is the probability that 5 cars pass and none are speeding? Assume that the speeds of the cars are independent of each other.

(b) On average, how many cars would the highway patrol officer expect to watch until the first car that is speeding? What is the standard deviation of the number of cars he would expect to watch?

4.49   Auto insurance premiums. Suppose a newspaper article states that the distribution of auto insurance premiums for residents of California is approximately normal with a mean of $1,650. The article also states that 25% of California residents pay more than $1,800.

(a) What is the Z-score that corresponds to the top 25% (or the 75th percentile) of the standard normal distribution?

(b) What is the mean insurance cost? What is the cutoff for the 75th percentile?

(c) Identify the standard deviation of insurance premiums in California.

4.50   SAT scores. SAT scores (out of 1600) are distributed normally with a mean of 1100 and a standard deviation of 200. Suppose a school council awards a certificate of excellence to all students who score at least 1350 on the SAT, and suppose we pick one of the recognized students at random. What is the probability this student's score will be at least 1500? (The material covered in Section 3.2 would be useful for this question.)

---

[36]S. Johnson and D. Murray. "Empirical Analysis of Truck and Automobile Speeds on Rural Interstates: Impact of Posted Speed Limits". In: *Transportation Research Board 89th Annual Meeting*. 2010.

4.51 Married women. The American Community Survey estimates that 47.1% of women ages 15 years and over are married.[37]

(a) We randomly select three women between these ages. What is the probability that the third woman selected is the only one who is married?

(b) What is the probability that all three randomly selected women are married?

(c) On average, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

(d) If the proportion of married women was actually 30%, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

(e) Based on your answers to parts (c) and (d), how does decreasing the probability of an event affect the mean and standard deviation of the wait time until success?

4.52 Survey response rate. Pew Research reported that the typical response rate to their surveys is only 9%. If for a particular survey 15,000 households are contacted, what is the probability that at least 1,500 will agree to respond?[38]

4.53 Overweight baggage. Suppose weights of the checked baggage of airline passengers follow a nearly normal distribution with mean 45 pounds and standard deviation 3.2 pounds. Most airlines charge a fee for baggage that weigh in excess of 50 pounds. Determine what percent of airline passengers incur this fee.

4.54 Heights of 10 year olds, Part I. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) What is the probability that a randomly chosen 10 year old is shorter than 48 inches?

(b) What is the probability that a randomly chosen 10 year old is between 60 and 65 inches?

(c) If the tallest 10% of the class is considered "very tall", what is the height cutoff for "very tall"?

4.55 Buying books on Ebay. The textbook you need to buy for your chemistry class is expensive at the college bookstore, so you consider buying it on Ebay instead. A look at past auctions suggest that the prices of that chemistry textbook have an approximately normal distribution with mean $89 and standard deviation $15.

(a) What is the probability that a randomly selected auction for this book closes at more than $100?

(b) Ebay allows you to set your maximum bid price so that if someone outbids you on an auction you can automatically outbid them, up to the maximum bid price you set. If you are only bidding on one auction, what are the advantages and disadvantages of setting a bid price too high or too low? What if you are bidding on multiple auctions?

(c) If you watched 10 auctions, roughly what percentile might you use for a maximum bid cutoff to be somewhat sure that you will win one of these ten auctions? Is it possible to find a cutoff point that will ensure that you win an auction?

(d) If you are willing to track up to ten auctions closely, about what price might you use as your maximum bid price if you want to be somewhat sure that you will buy one of these ten books?

4.56 Heights of 10 year olds, Part II. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) The height requirement for *Batman the Ride* at Six Flags Magic Mountain is 54 inches. What percent of 10 year olds cannot go on this ride?

(b) Suppose there are four 10 year olds. What is the chance that at least two of them will be able to ride *Batman the Ride*?

(c) Suppose you work at the park to help them better understand their customers' demographics, and you are counting people as they enter the park. What is the chance that the first 10 year old you see who can ride *Batman the Ride* is the 3rd 10 year old who enters the park?

(d) What is the chance that the fifth 10 year old you see who can ride *Batman the Ride* is the 12th 10 year old who enters the park?

---

[37]U.S. Census Bureau, 2010 American Community Survey, Marital Status.
[38]Pew Research Center, Assessing the Representativeness of Public Opinion Surveys, May 15, 2012.

4.57 Heights of 10 year olds, Part III. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) What fraction of 10 year olds are taller than 76 inches?

(b) If there are 2,000 10 year olds entering Six Flags Magic Mountain in a single day, then compute the expected number of 10 year olds who are at least 76 inches tall. (You may assume the heights of the 10-year olds are independent.)

(c) Using the binomial distribution, compute the probability that 0 of the 2,000 10 year olds will be at least 76 inches tall.

(d) The number of 10 year olds who enter Six Flags Magic Mountain and are at least 76 inches tall in a given day follows a Poisson distribution with mean equal to the value found in part (b). Use the Poisson distribution to identify the probability no 10 year old will enter the park who is 76 inches or taller.

4.58 Multiple choice quiz. In a multiple choice quiz there are 5 questions and 4 choices for each question (a, b, c, d). Robin has not studied for the quiz at all, and decides to randomly guess the answers. What is the probability that

(a) the first question she gets right is the 3rd question?

(b) she gets exactly 3 or exactly 4 questions right?

(c) she gets the majority of the questions right?

4.59 Overweight baggage, Part II. Suppose weights of the checked baggage of airline passengers follow a nearly normal distribution with mean 45 pounds and standard deviation 3.2 pounds. What is the probability that the *total* weight of 10 bags is greater than 460 lbs?

4.60 Chocolate chip cookies. Students are asked to count the number of chocolate chips in 22 cookies for a class activity. The packaging for these cookies claims that there are an average of 20 chocolate chips per cookie with a standard deviation of 4.37 chocolate chips.

(a) Based on this information, about how much variability should they expect to see in the mean number of chocolate chips in random samples of 22 chocolate chip cookies?

(b) What is the probability that a random sample of 22 cookies will have an average less than 14.77 chocolate chips if the company's claim on the packaging is true? Assume that the distribution of chocolate chips in these cookies is approximately normal.

(c) Assume the students got 14.77 as the average in their sample of 22 cookies. Do you have confidence or not in the company's claim that the true average is 20? Explain your reasoning.

4.61 Young Hispanics in the US. The 2012 Current Population Survey (CPS) estimates that 38.9% of the people of Hispanic origin in the Unites States are under 21 years old.[39] Calculate the probability that at least 35 people among a random sample of 100 Hispanic people living in the United States are under 21 years old.

4.62 Poverty in the US. The 2013 Current Population Survey (CPS) estimates that 22.5% of Mississippians live in poverty, which makes Mississippi the state with the highest poverty rate in the United States.[40] We are interested in finding out the probability that at least 250 people among a random sample of 1,000 Mississippians live in poverty.

(a) Estimate this probability using the normal approximation to the binomial distribution.

(b) Estimate this probability using the distribution of the sample proportion.

(c) How do your answers from parts (a) and (b) compare?

---

[39]United States Census Bureau.2012 Current Population Survey.The Hispanic Population in the United States: 2012. Web.

[40]United States Census Bureau. 2013 Current Population Survey.Historical Poverty Tables - People. Web.

Chapter 5

Foundations for inference

Statistical inference is primarily concerned with understanding and quantifying the uncertainty of parameter estimates. While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. We start with a familiar topic: the idea of using a sample proportion to estimate a population proportion. Next, we create what's called a *confidence interval*, which is a range of values where the true population value is likely to lie. Finally, we introduce a *hypothesis testing framework*, which allows us use data to formally evaluate claims about the population, such as whether a survey provides strong evidence that a candidate has the support of a majority of the voting population.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

5.1   Estimating unknown parameters

Companies such as the Gallup and Pew Research frequently conduct polls as a way to understand the state of public opinion or knowledge on many topics, including politics, scientific understanding, brand recognition, and more. How well do these polls estimate the opinion or knowledge of the broader population? Why is a larger sample generally preferable to a smaller sample? And what role does the concept of a sampling distribution, introduced in the previous chapter, play in answering these questions?

Learning objectives

1. Explain the difference between probability and inference and identify when to use which one.

2. Understand the purpose and use of a point estimate.

3. Understand how to measure the variability/error in a point estimate.

4. Recognize the relationship between the standard error of a point estimate and the standard deviation of a sample statistic.

5. Understand how changing the sample size affects the variability/error in a point estimate.

5.1.1 Point estimates

With this chapter, we move from the world of probability to the world of inference. Whereas **probability** involves using a known population value (parameter) to make a prediction about the likelihood of a particular sample value (statistic), **inference** involves using a calculated sample value (statistic) to estimate or better understand an unknown population value (parameter). For both of these, the concept of the sampling distribution is fundamental.

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a **point estimate** of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the **parameter** of interest, and when the parameter is a proportion, we denote it with the letter p. We typically estimate the parameter by collecting information from a sample of the population; we compute the observed proportion in the sample and we denote this sample proportion as $\hat{p}$. Unless we collect responses from every individual in the sample, p remains unknown, and we use $\hat{p}$ as our point estimate for p.

The difference we observe from the poll versus the parameter is called the **error** in the estimate. Generally, the error consists of two aspects: sampling error and bias.

Bias describes a systematic tendency to over- or under-estimate the true population value. For instance, if we took a political poll but our sample didn't include a roughly representative distribution of the political parties, the sample would likely skew in a particular direction and be biased. Taking a truly random sample helps avoid bias. However, as we saw in Chapter 1, even with a random sample, various types of response bias can still be present. For example, if we were taking a student poll asking about support for a new college stadium, we'd probably get a biased estimate of the stadium's level of student support by wording the question as, *Do you support your school by supporting funding for the new stadium?* We try to minimize bias through thoughtful data collection procedures, but bias can creep into our estimates without us even be aware.

Sampling error is uncertainty in a point estimate that happens naturally from one sample to the next. Much of statistics, including much of this book, is focused on understanding and quantifying sampling error. Remember though, that sampling error does not account for the possible effects of leading questions or other types of response bias. When we measure sampling error, we are measuring the expected variability in a point estimate that arises from randomly sampling only a subset of the population.

EXAMPLE 5.1

In Chapter 2, we found the summary statistics for the number of characters in a set of 50 email data. These values are summarized below.

$\bar{x}$	11,160
median	6,890
s_x	13,130

Estimate the **population mean** based on the sample.

The best estimate for the population mean is the **sample mean**. That is, $\bar{x} = 11,160$ is our best estimate for μ.

GUIDED PRACTICE 5.2

Using the email data, what quantity should we use as a point estimate for the population standard deviation σ?[1]

---

[1] Again, intuitively we would use the sample standard deviation $s = 13,130$ as our best estimate for σ.

5.1.2   Understanding the variability of a point estimate

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest.[2] If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%? We want to understand, *how does the sample proportion $\hat{p}$ behave when the true population proportion is 0.88*.[3] Let's find out! We can simulate responses we would get from a simple random sample of 1000 American adults, which is only possible because we know the actual support expanding solar energy to be 0.88. Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write "support" on 88% of them and "not" on the other 12%.

2. Mix up the pieces of paper and pull out 1000 pieces to represent our sample of 1000 American adults.

3. Compute the fraction of the sample that say "support".

Any volunteers to conduct this simulation? Probably not. While this physical simulation is totally impractical, we can simulate it thousands, even millions, of times using computer code. We've written a short computer simulation and run it 10,000 times. The results are show in Figure 5.1 in case you are curious what the computer code looks like. In this simulation, the sample gave a point estimate of $\hat{p}_1 = 0.894$. We know the population proportion for the simulation was $p = 0.88$, so we know the estimate had an error of $0.894 - 0.88 = +0.014$.

```
# 1. Create a set of 250 million entries, where 88% of them are "support"
#    and 12% are "not".
pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size), rep("not", 0.12 * pop_size))

# 2. Sample 1000 entries without replacement.
sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support", then divide by
#    the sample size.
sum(sampled_entries == "support") / 1000
```

Figure 5.1: For those curious, this is code for a single $\hat{p}$ simulation using the statistical software called **R**. Each line that starts with **#** is a **code comment**, which is used to describe in regular language what the code is doing. We've provided software labs in **R** at openintro.org/stat/labs for anyone interested in learning more.

---

[2]We haven't actually conducted a census to measure this value perfectly. However, a very large sample has suggested the actual level of support is about 88%.

[3]Note: 88% written as a proportion would be 0.88. It is common to switch between proportion and percent.

One simulation isn't enough to get a great sense of the distribution of estimates we might expect in the simulation, so we should run more simulations. In a second simulation, we get $\hat{p}_2 = 0.885$, which has an error of +0.005. In another, $\hat{p}_3 = 0.878$ for an error of -0.002. And in another, an estimate of $\hat{p}_4 = 0.859$ with an error of -0.021. With the help of a computer, we've run the simulation 10,000 times and created a histogram of the results from all 10,000 simulations in Figure 5.2. This distribution of sample proportions is called a **sampling distribution**. We can characterize this sampling distribution as follows:

Center. The center of the distribution is $\mu_{\hat{p}} = 0.880$, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straight-forward sampling strategy that helps avoid sampling bias.

Spread. The standard deviation of the distribution is $\sigma_{\hat{p}} = 0.010$.

Shape. The distribution is symmetric and bell-shaped, and it *resembles a normal distribution.*

These findings are encouraging! When the population proportion is $p = 0.88$ and the sample size is $n = 1000$, the sample proportion $\hat{p}$ tends to give a pretty good estimate of the population proportion. We also have the interesting observation that the histogram resembles a normal distribution.
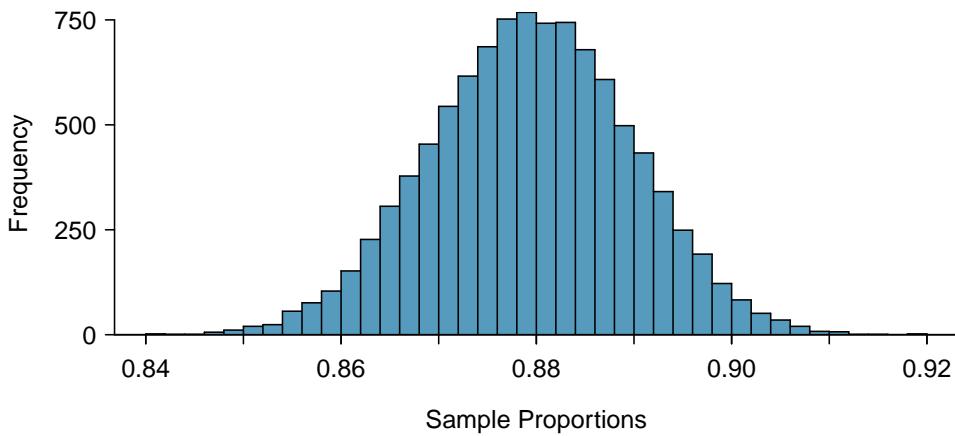
Figure 5.2: A histogram of 10,000 sample proportions sampled from a population where the population proportion is 0.88 and the sample size is $n = 1000$.

SAMPLING DISTRIBUTIONS ARE NEVER OBSERVED, BUT WE KEEP THEM IN MIND

In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution. Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

EXAMPLE 5.3

If we used a much smaller sample size of $n = 50$, would you guess that the standard error for $\hat{p}$ would be larger or smaller than when we used $n = 1000$?

Intuitively, it seems like more data is better than less data, and generally that is correct! The typical error when $p = 0.88$ and $n = 50$ would be larger than the error we would expect when $n = 1000$.

Example 5.3 highlights an important property we will see again and again: a bigger sample tends to provide a more precise point estimate than a smaller sample. Remember though, that this is only true for *random* samples. Additionally, a bigger sample cannot correct for response bias or other types of bias that may be present.

5.1.3   Introducing the standard error

Point estimates only approximate the population parameter. How can we quantify the expected variability in a point estimate $\hat{p}$? The discussion in Section 4.5 tells us how. The variability in the distribution of $\hat{p}$ is given by its standard deviation. If we know the population proportion, we can calculate the standard deviation of the point estimate $\hat{p}$. In our simulation we knew p was 0.88. Thus we can calculate the standard deviation as

$$SD_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.088)(1-0.88)}{n}} = 0.01$$

If we now look at the sampling distribution, we see that the typical distance sample proportions are from the true value of 0.88 is about 0.01.

EXAMPLE 5.4

Consider a random sample of size 80 from a population. We find that 15% of the sample support a controversial new ballot measure. How far is our estimate likely to be from the true percent that support the measure?

Ⓔ

We would like to calculate the standard deviation of $\hat{p}$, but we run into a serious problem: p is *unknown*. In fact, when doing inference, p must be unknown, otherwise it is illogical to try to estimate it. We cannot calculate the SD, but we can estimate it using, you might have guessed, the sample proportion $\hat{p}$.

This estimate of the standard deviation is known as the **standard error**, or **SE** for short.

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

EXAMPLE 5.5

Calculate and interpret the SE of $\hat{p}$ for the previous example.

Ⓔ

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.15(1-0.15)}{80}} = 0.04$$

The typical or expected error in our estimate is 4%.

EXAMPLE 5.6

If we quadruple the sample size from 80 to 320, what will happen to the SE?

Ⓔ

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.15(1-0.15)}{320}} = 0.02$$

The larger the sample size, the smaller our standard error. This is consistent with intuition: the more data we have, the more reliable an estimate will tend to be. However, quadrupling the sample size does not reduce the error by a factor of 4. Because of the square root, the effect is to reduce the error by a factor $\sqrt{4}$, or 2.

5.1.4 Basic properties of point estimates

We achieved three goals in this section. First, we determined that point estimates from a sample may be used to estimate population parameters. We also determined that these point estimates are not exact: they vary from one sample to another. Lastly, we quantified the uncertainty of the sample proportion using what we call the standard error.

Remember that the standard error only measures sampling error. It does not account for bias that results from leading questions or other types of response bias.

When our sampling method produces estimates in an unbiased way, the sampling distribution will be *centered* on the true value and we call the method **accurate**. When the sampling method produces estimates that have *low variability*, the sampling distribution will have a low standard error, and we call the method **precise**.
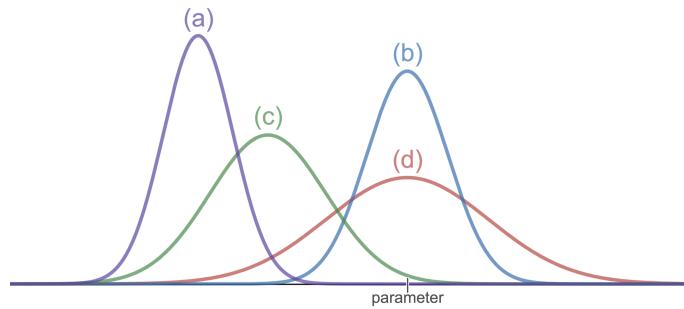
Figure 5.3: Four sampling distributions shown, with parameter identified. Explore these distributions through a Desmos activity at openintro.org/ahss/desmos.

EXAMPLE 5.7

Using Figure 5.3, which of the distributions were produced by methods that are biased? That are accurate? Order the distributions from most precise to least precise (that is, from lowest variability to highest variability).

———

Distributions (b) and (d) are centered on the parameter (the true value), so those methods are accurate. The methods that produced distributions (a) and (c) are biased, because those distributions are not centered on the parameter. From most precise to least precise, we have (a), (b), (c), (d).

EXAMPLE 5.8

Why do we want a point estimate to be both precise and accurate?

———

If the point estimate is precise, but highly biased, then we will consistently get a bad estimate. On the other hand, if the point estimate is unbiased but not at all precise, then by random chance, we may get an estimate far from the true value.

Remember, when taking a sample, we generally get only one chance. It is the properties of the sampling distribution that tell us how much confidence we can have in the estimate.

The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion. For instance, if we want to estimate the average salary for graduates from a particular college, we could survey a random sample of recent graduates; in that example, we'd be using a sample mean $\bar{x}$ to estimate the population mean μ for all graduates. As another example, if we want to estimate the difference in product prices for two websites, we might take a random sample of products available on both sites, check the prices on each, and use then compute the average difference; this strategy certainly wouldn't give us a perfect measurement of the actual difference, but it would give us a point estimate.

While this chapter emphases a single proportion context, we'll encounter many different contexts throughout this book where these methods will be applied. The principles and general ideas are the same, even if the details change a little.

Section summary

- In this section we laid the groundwork for our study of **inference**. Inference involves using known sample values to estimate or better understand unknown population values.

- A sample statistic can serve as a **point estimate** for an unknown parameter. For example, the sample mean is a point estimate for an unknown population mean, and the sample proportion is a point estimate for an unknown population proportion.

- It is helpful to imagine a point estimate as being drawn from a particular sampling distribution.

- The **standard error** (**SE**) of a point estimate tells us the typical error or uncertainty associated with the point estimate. It is also an estimate of the spread of the sampling distribution.

- A point estimate is **unbiased** (accurate) if the sampling distribution (i.e., the distribution of all possible outcomes of the point estimate from repeated samples from the same population) is *centered* on the true population parameter.

- A point estimate has **lower variability** (more precise) when the *standard deviation* of the sampling distribution is smaller.

  item In a random sample, increasing the sample size n will make the standard error smaller. This is consistent with the intuition that larger samples tend to be more reliable, all other things being equal.

- In general, we want a point estimate to be unbiased and to have low variability. Remember: the terms unbiased (accurate) and low variability (precise) are properties of generic point estimates, which are variables that have a *sampling distribution*. These terms do not apply to individual values of a point estimate, which are *numbers*.

Exercises

5.1 Identify the parameter, Part I. For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(a) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

(b) In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

(c) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

(d) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

(e) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

5.2 Identify the parameter, Part II. For each of the following situations, state whether the parameter of interest is a mean or a proportion.

(a) A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.

(b) A survey reports that local TV news has shown a 17% increase in revenue within a two year period while newspaper revenues decreased by 6.4% during this time period.

(c) In a survey, high school and college students are asked whether or not they use geolocation services on their smart phones.

(d) In a survey, smart phone users are asked whether or not they use a web-based taxi service.

(e) In a survey, smart phone users are asked how many times they used a web-based taxi service over the last year.

5.3 Quality control. As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) What is the point estimate for the parameter?

(d) What is the name of the statistic can we use to measure the uncertainty of the point estimate?

(e) Compute the value from part (d) for this context.

(f) The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

(g) Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (e) using $p = 0.1$ instead of $\hat{p}$, does the resulting value change much?

5.4 Unexpected expense. In a random sample 765 adults in the United States, 322 say they could not cover a $400 unexpected expense without borrowing money or going into debt.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) What is the point estimate for the parameter?

(d) What is the name of the statistic can we use to measure the uncertainty of the point estimate?

(e) Compute the value from part (d) for this context.

(f) A cable news pundit thinks the value is actually 50%. Should she be surprised by the data?

(g) Suppose the true population value was found to be 40%. If we use this proportion to recompute the value in part (e) using $p = 0.4$ instead of $\hat{p}$, does the resulting value change much?

5.5   Repeated water samples. A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) If the proportions are distributed around 8%, what is the variability of the distribution?

(d) What is the formal name of the value you computed in (c)?

(e) Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

5.6   Repeated student samples. Of all freshman at a large college, 16% made the dean's list in the current year. As part of a class project, students randomly sample 40 students and check if those students made the list. They repeat this 1,000 times and build a distribution of sample proportions.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) Calculate the variability of this distribution.

(d) What is the formal name of the value you computed in (c)?

(e) Suppose the students decide to sample again, this time collecting 90 students per sample, and they again collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 40 observations?

5.2 Confidence intervals

The site fivethirtyeight.com regularly forecasts support for each candidate in Congressional races, i.e. races in the US House of Representatives and the US Senate. In addition to point estimates, they report confidence intervals.[4] What are confidence intervals, and how do we interpret them?

Learning objectives

1. Explain the purpose and use of confidence intervals.

2. Construct 95% confidence intervals assuming the point estimate follows a normal distribution.

3. Calculate the critical value for a C% confidence interval when the point estimate follows a normal distribution.

4. Describe how sample size and confidence level affect the width of a confidence interval.

5. Interpret a confidence interval and the confidence level in context.

6. Draw conclusions with a specified confidence level about the values of unknown parameters.

7. Calculate and interpret the margin of error for a C% confidence interval. Distinguish between margin of error and standard error.

5.2.1 Capturing the population parameter

A point estimate provides a single plausible value for a parameter. However, a point estimate isn't perfect and will have some *standard error* associated with it. When estimating a parameter, it is better practice to provide a plausible *range of values* instead of supplying just the point estimate.

A plausible range of values for the population parameter is called a **confidence interval**. Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

5.2.2   Constructing a 95% confidence interval

A point estimate is our best guess for the value of the parameter, so it makes sense to build the confidence interval around that value. The standard error is a measure of the uncertainty associated with the point estimate.

EXAMPLE 5.9

How many standard errors should we extend above and below the point estimate if we want to be 95% confident of capturing the true value?

First, we observe that the area under the standard normal curve between -1.96 and 1.96 is 95%. When conditions for a normal model are met, the point estimate we observe will be within 1.96 standard deviations of the true value about 95% of the time. Thus, if we want to be 95% confident of capturing the true value, we should go 1.96 standard errors on either side of the point estimate.

CONSTRUCTING A 95% CONFIDENCE INTERVAL USING A NORMAL MODEL

When the sampling distribution of a point estimate can reasonably be modeled as normal, a 95% confidence interval for the unknown parameter can be constructed as:

$$\text{point estimate} \; \pm \; 1.96 \times SE \text{ of estimate} \qquad (5.10)$$

We can be **95% confident** that this interval captures the true value.

In the next chapters we will determine when we can apply a normal model to a point estimate. For now, we will assume that a normal model is reasonable.

EXAMPLE 5.11

The point estimate from the smoking example was 15%. The standard error for this point estimate was calculated to be $SE = 0.04$. Assuming that conditions for a normal model are met, construct and interpret a 95% confidence interval.

$$\text{point estimate} \; \pm \; 1.96 \times SE \text{ of estimate}$$
$$0.15 \; \pm \; 1.96 \times 0.04$$
$$(0.0716,\ 0.2284)$$

We are 95% confident that the true percent of smokers in this population is between 7.16% and 22.84%.

EXAMPLE 5.12

Based on the confidence interval above, is there evidence that a smaller proportion smoke in this county than in the state as a whole? The proportion that smoke in the state is known to be 0.20.

While the point estimate of 0.15 is lower than 0.20, this deviation is likely due to random chance. Because the confidence interval *includes* the value 0.20, 0.20 is a reasonable value for the proportion of smokers in the county. Therefore, based on this confidence interval, we do not have evidence that a smaller proportion smoke in the county than in the state.

We can be 95% confident that a 95% confidence interval captures the true population parameter. However, confidence intervals are imperfect. About 1-in-20 (5%) properly constructed 95% confidence intervals will fail to capture the parameter of interest. Figure 5.4 shows 25 confidence intervals for a proportion that were constructed from simulations where the true proportion was $p = 0.3$. However, 1 of these 25 confidence intervals happened not to include the true value.
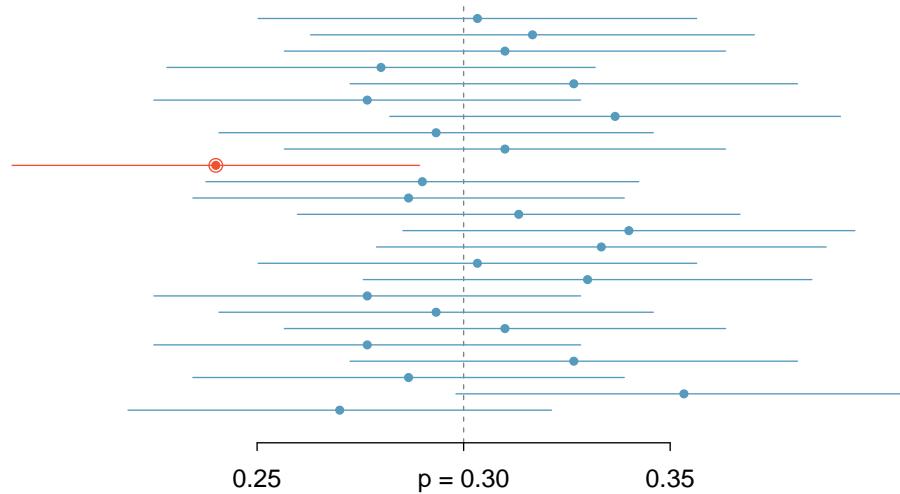
Figure 5.4: Twenty-five samples of size $n = 300$ were simulated when $p = 0.30$. For each sample, a confidence interval was created to try to capture the true proportion p. However, 1 of these 25 intervals did not capture $p = 0.30$.

GUIDED PRACTICE 5.13

In Figure 5.4, one interval does not capture the true proportion, $p = 0.3$. Does this imply that there was a problem with the simulations?[5]

5.2.3 Changing the confidence level

Suppose we want to construct a confidence interval with a confidence level somewhat greater than 95%: perhaps we would like a confidence level of 99%.

EXAMPLE 5.14

Other things being equal, would a 99% confidence interval be wider or narrower than a 95% confidence interval?

———

Using a previous analogy: if we want to be more confident that we will catch a fish, we should use a wider net, not a smaller one. To be 99% confidence of capturing the true value, we must use a wider interval. On the other hand, if we want an interval with lower confidence, such as 90%, we would use a narrower interval.

The 95% confidence interval structure provides guidance in how to make intervals with new confidence levels. Below is a general 95% confidence interval for a point estimate that comes from a nearly normal distribution:

$$\text{point estimate} \ \pm \ 1.96 \times SE \text{ of estimate} \tag{5.15}$$

There are three components to this interval: the point estimate, "1.96", and the standard error. The choice of $1.96 \times SE$ was based on capturing 95% of the distribution since the estimate is within 1.96 standard deviations of the true value about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

---

[5]No. Just as some observations occur more than 1.96 standard deviations from the mean, some point estimates will be more than 1.96 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are impossible.

GUIDED PRACTICE 5.16

Ⓖ

If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[6]
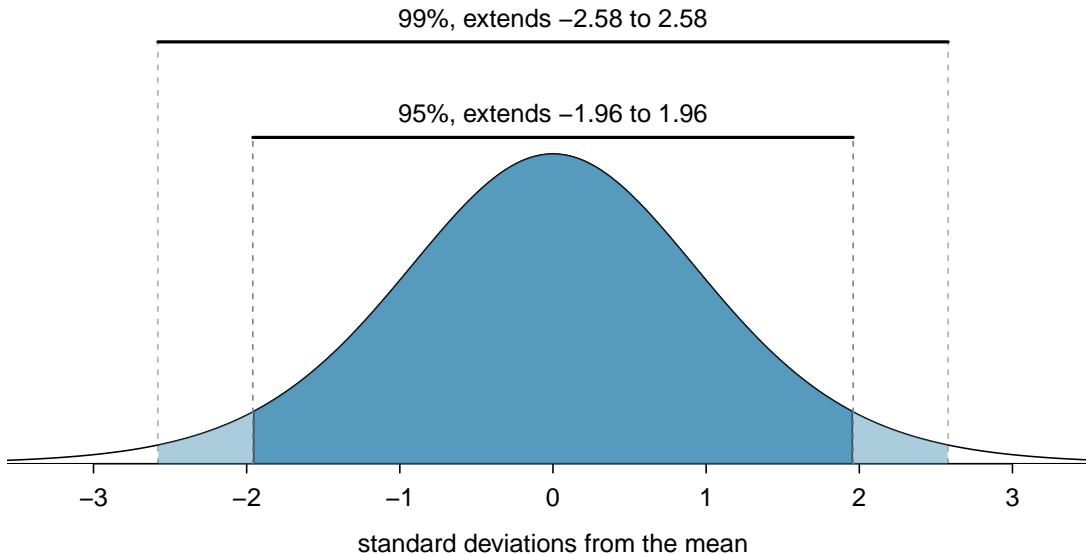
Figure 5.5: The area between $-z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between $-z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58. Guided Practice 5.16 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of its mean. Thus, the formula for a 99% confidence interval is

$$\text{point estimate} \; \pm \; 2.58 \times SE \text{ of estimate} \tag{5.17}$$

Figure 5.5 provides a picture of how to identify $z^\star$ based on a confidence level.

The *number* of standard errors we go above and below the point estimate is called the **critical value**. When the critical value is determined based on a normal model, we call the critical value $z^\star$.

CONFIDENCE INTERVAL FOR ANY CONFIDENCE LEVEL

If the point estimate follows a normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \; \pm \; z^\star \times SE \text{ of estimate}$$

where $z^\star$ depends on the confidence level selected.

Finding the value of $z^\star$ that corresponds to a particular confidence level is most easily accomplished by using a new table, called the t-table. For now, what is noteworthy about this table is that the bottom row corresponds to confidence levels. The numbers inside the table are the critical values, but which row should we use? Later in this book, we will see that a t-curve with infinite degrees of freedom corresponds to the normal curve. For this reason, when finding $z^\star$, we use the t-table at row ∞.

---

[6]This is equivalent to asking how often the Z-score will be larger than -2.58 but less than 2.58. (For a picture, see Figure 5.5.) There is ≈ 0.99 probability that the unobserved random variable X will be within 2.58 standard deviations of the mean.

	one tail	0.100	0.050	0.025	0.010	0.005
df	1	3.078	6.314	12.71	31.82	63.66
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	1000	1.282	1.646	1.962	2.330	2.581
	∞	1.282	1.645	1.960	2.326	2.576
Confidence level C		80%	90%	95%	98%	99%

Figure 5.6: An abbreviated look at the *t*-table. The columns correspond to confidence levels. Row ∞ corresponds to the normal curve.

FINDING $z^\star$ FOR A PARTICULAR CONFIDENCE LEVEL

We select $z^\star$ so that the area between $-z^\star$ and $z^\star$ in the normal model corresponds to the confidence level. Use a calculator or use the *t*-table at row ∞ to find the critical value $z^\star$.

GUIDED PRACTICE 5.18

Find the appropriate $z^\star$ value for an 80% confidence interval.[7]

The normal approximation is crucial to the precision of these confidence intervals. The next two chapters provide detailed discussions about when a normal model can safely be applied to a variety of situations. When a normal model is not a good fit, we will use alternate distributions that better characterize the sampling distribution.

5.2.4 Margin of error

The confidence intervals we have encountered thus far have taken the form

$$\text{point estimate } \pm \ z^* \times SE \text{ of estimate}$$

Confidence intervals are also often reported as

$$\text{point estimate } \pm \ \text{margin of error}$$

For example, instead of reporting an interval as $0.09 \pm 1.645 \times 0.028$ or $(0.044, 0.136)$, it could be reported as 0.09 ± 0.046.

The **margin of error** is the distance between the point estimate and the lower or upper bound of a confidence interval. It is half of the total width of the interval.

MARGIN OF ERROR

When the point estimate follows a normal distribution,

$$\text{margin of error} = z^\star \times SE \text{ of estimate.}$$

---

[7]Using row ∞ on the *t*-table, we see that the value that corresponds to an 80% confidence level is 1.282. Therefore, we should use 1.282 as the $z^\star$ value.

EXAMPLE 5.19

All other things being equal, will the margin of error be bigger for a 68% confidence interval or a 95% confidence interval?

A 95% confidence interval is wider than a 68% confidence interval and has a larger $z^\star$ value, so the 95% confidence interval will have a larger margin of error.

GUIDED PRACTICE 5.20

What is the margin of error for the confidence interval: (0.035, 0.145)?[8]

5.2.5 Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

We are C% confident that the population parameter is between ___ and ___.

Incorrect language might try to describe the confidence interval as capturing the population parameter with a certain probability.[9] Applying the language of probability to a fixed interval or to a fixed parameter is one of the most common errors.

As we saw in Figure 5.4, the 95% confidence interval *method* has a 95% probability of producing an interval that will capture the population parameter. A correct interpretation of the confidence *level* is that such intervals will capture the population parameter that percent of the time (assuming conditions are met and the probability model is true). However, each individual interval either does or does not capture the population parameter. A correct interpretation of an individual confidence interval cannot involve the vocabulary of probability.

Another especially important consideration of confidence intervals is that they *only try to capture the population parameter.* Our intervals say nothing about the confidence of capturing individual observations, a proportion of the observations, or point estimates. Confidence intervals only attempt to capture population parameters.

5.2.6 Confidence interval procedures: a five step process

Use a confidence interval to *estimate* a parameter with a particular *confidence level*, C.

(AP EXAM TIP) WHEN CARRYING OUT A CONFIDENCE INTERVAL PROCEDURE, FOLLOW THESE FIVE STEPS:

- **Identify**: Identify the parameter and the confidence level.
- **Choose**: Choose the appropriate interval procedure and identify it by name.
- **Check**: Check that the conditions for the interval procedure are met.
- **Calculate**: Calculate the confidence interval and record it in interval form.

$$\text{CI: point estimate } \pm \text{ critical value} \times SE \text{ of estimate}$$

- **Conclude**: Interpret the interval and, if applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value of interest.

---

[8]The margin of error is *half* of the total width of the interval. The margin of error for this interval is $\frac{0.145-0.035}{2} = 0.055$.

[9]To see that this interpretation is incorrect, imagine taking two random samples and constructing two 95% confidence intervals for an unknown proportion. If these intervals are disjoint, can we say that there is a 95%+95%=190% chance that the first or the second interval captures the true value?

Section summary

- A point estimate is not perfect; there is almost always some error in the estimate. It is often useful to supply a plausible *range of values* for the parameter, which we call a **confidence interval**.

- A confidence interval is centered on the point estimate and extends a certain number of standard errors on either side of the estimate, depending upon how *confident* one wants to be. For a fixed sample size, to be more confident of capturing the true value requires a wider interval.

- When the sampling distribution of a point estimate can reasonably be modeled as *normal*, such as with a **sample proportion**, then the following are true:

    - A 68% confidence interval is given by: point estimate $\pm$ SE of estimate.
      We can be 68% confident this interval captures the true value.
    - A 95% confidence interval is given by: point estimate $\pm$ $1.96 \times SE$ of estimate.
      We can be 95% confident this interval captures the true value.
    - A C% confidence interval is given by: point estimate $\pm$ $z^\star \times SE$ of estimate.
      We can be C% confident this interval captures the true value.

- For a C% confidence interval described above, we select $z^\star$ such that the area between $-z^\star$ and $z^\star$ under the standard normal curve is C%. Use the *t*-table at row ∞ to find the critical value $z^\star$.[10]

- After interpreting the interval, we can usually draw a conclusion, with C% confidence, about whether a given value X is a reasonable value for the population parameter. When drawing a conclusion based on a confidence interval, there are three possibilities.

    - We *have evidence* that the true [parameter]:
          ...is greater than X, because the entire interval is *above* X.
          ...is less than X, because the entire interval is *below* X.
    - We *do not have evidence* that the true [parameter] is not X, because X is *in* the interval.

- AP exam tip: A full confidence interval procedure includes the following steps.

    1. **Identify**: Identify the parameter and the confidence level.
    2. **Choose**: Choose the appropriate interval procedure and identify it by name.
    3. **Check**: Check that the conditions for the interval procedure are met.
    4. **Calculate**: Calculate the confidence interval and record it in interval form.

        CI: point estimate $\pm$ critical value $\times SE$ of estimate

    5. **Conclude**: Interpret the interval and, if applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value of interest.

Interpreting **confidence intervals** and **confidence levels**

- 68% and 95% are examples of **confidence levels**. The confidence level tells us the *capture rate* with repeated sampling. For example, a correct interpretation of a 95% confidence level is that if many samples of the same size were taken from the population, about 95% of the resulting confidence intervals would capture the true population parameter (assuming the conditions are met and the probability model is true). Note that this is a *relative frequency interpretation*.

- We cannot use the language of probability to interpret an *individual* confidence interval, once it has been calculated. The confidence level tells us what percent of the intervals will capture the population parameter, not the probability that a calculated interval captures the population parameter. Each calculated interval either does or does not capture the population parameter.

---

[10]We explain the relationship between z and t in the next chapter.

Margin of error

- Confidence intervals are often reported as: point estimate $\pm$ margin of error. The **margin of error** (ME) = critical value $\times$ SE of estimate, and it tells us, with a particular confidence, how much we expect our point estimate to deviate from the true population value due to chance.

- The margin of error depends on the *confidence level*; the standard error does not. Other things being constant, a higher confidence level leads to a larger margin of error.

- For a fixed confidence level, increasing the sample size decreases the margin of error. This assumes a random sample.

- The margin of error formula only applies if a sample is random. Moreover, the margin of error measures only *sampling error*; it does not account for additional error introduced by response bias and non-response bias. Even with a perfectly random sample, the actual error in a poll is likely higher than the reported margin of error.[11]

---

[11]nytimes.com/2016/10/06/upshot/when-you-hear-the-margin-of-error-is-plus-or-minus-3-percent-think-7-instead.html

Exercises

5.7   Chronic illness, Part I. In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions".[12] However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

5.8   Twitter users and news, Part I. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter.[13]. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

5.9   Chronic illness, Part II. In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions", and the standard error for this estimate is 1.2%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) We can say with certainty that the confidence interval from Exercise 5.7 contains the true percentage of U.S. adults who suffer from a chronic illness.

(b) If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

(c) The poll provides statistically significant evidence (at the $\alpha = 0.05$ level) that the percentage of U.S. adults who suffer from chronic illnesses is below 50%.

(d) Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

5.10   Twitter users and news, Part II. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter, and the standard error for this estimate was 2.4%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) The data provide statistically significant evidence that more than half of U.S. adult Twitter users get some news through Twitter. Use a significance level of $\alpha = 0.01$.

(b) Since the standard error is 2.4%, we can conclude that 97.6% of all U.S. adult Twitter users were included in the study.

(c) If we want to reduce the standard error of the estimate, we should collect less data.

(d) If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than a corresponding 99% confidence interval.

---

[12]Pew Research Center, Washington, D.C. The Diagnosis Difference, November 26, 2013.
[13]Pew Research Center, Washington, D.C. Twitter News Consumers: Young, Mobile and Educated, November 4, 2013.

5.11   Waiting at an ER, Part I. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning.

(a) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

(b) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.

(c) 95% of random samples have a sample mean between 128 and 147 minutes.

(d) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

(e) The margin of error is 9.5 and the sample mean is 137.5.

(f) In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size.

5.12   Mental health. The General Social Survey asked the question: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

(a) Interpret this interval in context of the data.

(b) What does "95% confident" mean? Explain in the context of the application.

(c) Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or wider than the 95% confidence interval?

(d) If a new survey were to be done with 500 Americans, do you think the standard error of the estimate be larger, smaller, or about the same.

5.3 Introducing hypothesis testing

In an experiment, one treatment reduces cholesterol by 10% while another treatment reduces it by 17%. Is this strong enough evidence that the second treatment is more effective? In this section, we will set up a framework for answering questions such as this and will look at the different types of decision errors that researcher can make when drawing conclusions based on data.

Learning objectives

1. Explain the logic of hypothesis testing, including setting up hypotheses and drawing a conclusion based on the set significance level and the calculated p-value.

2. Set up the null and alternative hypothesis in words and in terms of population parameters.

3. Interpret a p-value in context and recognize how the calculation of the p-value depends upon the direction of the alternative hypothesis.

4. Define and interpret the concept statistically significant.

5. Interpret Type I, Type II Error, and power in the context of hypothesis testing.

5.3.1 Case study: medical consultant

People providing an organ for donation sometimes seek the help of a special medical consultant. These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients.

One consultant tried to attract patients by noting the overall complication rate for liver donor surgeries in the US is about 10%, but her clients have had only 9 complications in the 142 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

EXAMPLE 5.21

We will let p represent the true complication rate for liver donors working with this consultant. Calculate the best estimate for p using the data. Label the point estimate as $\hat{p}$.

The sample proportion for the complication rate is 9 complications divided by the 142 surgeries the consultant has worked on: $\hat{p} = 9/142 = 0.063$.

EXAMPLE 5.22

Is it possible to prove that the consultant's work reduces complications?

No. The claim implies that there is a causal connection, but the data are observational. For example, maybe patients who can afford a medical consultant can afford better medical care, which can also lead to a lower complication rate.

EXAMPLE 5.23

While it is not possible to assess the causal claim, it is still possible to ask whether the low complication rate of $\hat{p} = 0.063$ provides evidence that the consultant's true complication rate is different than the US complication rate. Why might we be tempted to immediately conclude that the consultant's true complication rate is different than the US complication rate? Can we draw this conclusion?

___

Her sample complication rate is $\hat{p} = 0.063$, which is 0.037 lower than the US complication rate of 10%. However, we cannot yet be sure if the observed difference represents a real difference or is just the result of random variation. We wouldn't expect the sample proportion to be *exactly* 0.10, even if the truth was that her real complication rate was 0.10.

5.3.2   Setting up the null and alternate hypothesis

We can set up two competing hypotheses about the consultant's true complication rate. The first is call the **null hypothesis** and represents either a skeptical perspective or a perspective of no difference. The second is called the **alternative hypothesis** (or alternate hypothesis) and represents a new perspective such as the possibility that there has been a change or that there is a treatment effect in an experiment.

NULL AND ALTERNATIVE HYPOTHESES

The **null hypothesis** is abbreviated H_0. It represents a skeptical perspective and is often a claim of no change or no difference.

The **alternative hypothesis** is abbreviated H_A. It is the claim researchers hope to prove or find evidence for, and it often asserts that there has been a change or an effect.

Our job as data scientists is to play the skeptic: before we buy into the alternative hypothesis, we need to see strong supporting evidence.

EXAMPLE 5.24

Identify the null and alternative claim regarding the consultant's complication rate.

___

H_0: The true complication rate for the consultant's clients is the *same as* the US complication rate of 10%.

H_A: The true complication rate for the consultant's clients is different than 10%.

Often it is convenient to write the null and alternative hypothesis in mathematical or numerical terms. To do so, we must first identify the quantity of interest. This quantity of interest is known as the parameter for a hypothesis test.

PARAMETERS AND POINT ESTIMATES

A **parameter** for a hypothesis test is the "true" value of the population of interest. When the parameter is a proportion, we call it p.

A **point estimate** is calculated from a sample. When the point estimate is a proportion, we call it $\hat{p}$.

The observed or sample proportion of 0.063 is a point estimate for the true proportion. The parameter in this problem is the true proportion of complications for this consultant's clients. The parameter is unknown, but the null hypothesis is that it equals the overall proportion of complications: $p = 0.10$. This hypothesized value is called the null value.

NULL VALUE OF A HYPOTHESIS TEST

The **null value** is the value hypothesized for the parameter in H_0, and it is sometimes represented with a subscript 0, e.g. p_0 (just like H_0).

In the medical consultant case study, the parameter is p and the null value is $p_0 = 0.10$. We can write the null and alternative hypothesis as numerical statements as follows.

- H_0: $p = 0.10$ (The complication rate for the consultant's clients is equal to the US complication rate of 10%.)

- H_A: $p \neq 0.10$ (The complication rate for the consultant's clients is not equal to the US complication rate of 10%.)

HYPOTHESIS TESTING

These hypotheses are part of what is called a **hypothesis test**. A hypothesis test is a statistical technique used to evaluate competing claims using data. Often times, the null hypothesis takes a stance of *no difference* or *no effect*. If the null hypothesis and the data notably disagree, then we will reject the null hypothesis in favor of the alternative hypothesis.

Don't worry if you aren't a master of hypothesis testing at the end of this section. We'll discuss these ideas and details many times in this chapter and the two chapters that follow.

The null claim is always framed as an equality: it tells us what quantity we should use for the parameter when carrying out calculations for the hypothesis test. There are three choices for the alternative hypothesis, depending upon whether the researcher is trying to prove that the value of the parameter is greater than, less than, or not equal to the null value.

ALWAYS WRITE THE NULL HYPOTHESIS AS AN EQUALITY

We will find it most useful if we always list the null hypothesis as an equality (e.g. $p = 7$) while the alternative always uses an inequality (e.g. $p \neq 0.7$, $p > 0.7$, or $p < 0.7$).

GUIDED PRACTICE 5.25

According to the 2010 US Census, 7.6% of residents in the state of Alaska were under 5 years old. A researcher plans to take a random sample of residents from Alaska to test whether or not this is still the case. Write out the hypotheses that the researcher should test in both plain and statistical language.[14]

When the alternative claim uses a $\neq$, we call the test a **two-sided** test, because either extreme provides evidence against H_0. When the alternative claim uses a $<$ or a $>$, we call it a **one-sided** test.

ONE-SIDED AND TWO-SIDED TESTS

If the researchers are only interested in showing an increase or a decrease, but not both, use a one-sided test. If the researchers would be interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

---

[14]H_0: $p = 0.076$; The proportion of residents under 5 years old in Alaska is *unchanged* from 2010. H_A: $p \neq 0.076$; The proportion of residents under 5 years old in Alaska has changed from 2010. Note that it could have increased or decreased. `https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml`

EXAMPLE 5.26

For the example of the consultant's complication rate, we knew that her sample complication rate was 0.063, which was lower than the US complication rate of 0.10. Why did we conduct a two-sided hypothesis test for this setting?

———————

The setting was framed in the context of the consultant being helpful, but what if the consultant actually performed worse than the US complication rate? Would we care? More than ever! Since we care about a finding in either direction, we should run a two-sided test.

ONE-SIDED HYPOTHESES ARE ALLOWED ONLY *BEFORE* SEEING DATA

After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Hypotheses must be set up *before* observing the data. If they are not, the test must be two-sided.

5.3.3   Evaluating the hypotheses with a p-value

EXAMPLE 5.27

There were 142 patients in the consultant's sample. If the null claim is true, how many would we expect to have had a complication?

———————

If the null claim is true, we would expect about 10% of the patients, or about 14.2 to have a complication.

The consultant's complication rate for her 142 clients was 0.063 ($0.063 \times 142 \approx 9$). What is the probability that a sample would produce a number of complications this far from the expected value of 14.2, *if her true complication rate were* 0.10, that is, if H_0 were true? The probability, which is estimated in Section 5.7 on page 282, is about 0.1754. We call this quantity the **p-value**.
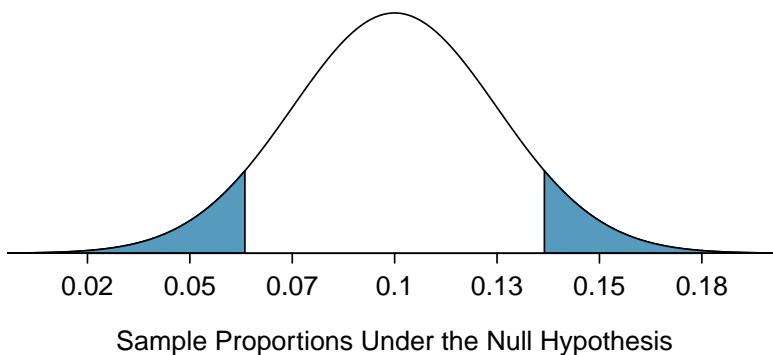
Figure 5.7: The shaded area represents the p-value. We observed $\hat{p} = 0.063$, so any observations smaller than this are at least as extreme relative to the null value, $p_0 = 0.1$, and so the lower tail is shaded. However, since this is a two-sided test, values above 0.137 are also at least as extreme as 0.063 (relative to 0.1), and so they also contribute to the p-value. The tail areas together total of about 0.1754 when calculated using a simulation technique in Section 5.3.4.
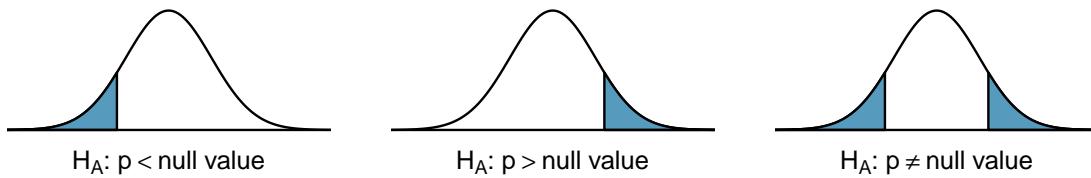
Figure 5.8: When the alternative hypothesis takes the form $p <$ null value, the p-value is represented by the lower tail. When it takes the form $p >$ null value, the p-value is represented by the upper tail. When using $p \neq$ null value, then the p-value is represented by both tails.

FINDING AND INTERPRETING THE P-VALUE

We find and interpret the **p-value** according to the nature of the alternative hypothesis.

H_A: **parameter > null value.** The p-value corresponds to the area in the *upper* tail and is probability of getting a test statistic larger than the observed test statistic if the null hypothesis is true and the probability model is accurate.

H_A: **parameter < null value.** The p-value corresponds to the area in the *lower* tail and is the probability of observing a test statistic smaller than the observed test statistic if the null hypothesis is true and the probability model is accurate.

H_A: **parameter $\neq$ null value.** The p-value corresponds to the area in *both* tails and is the probability of observing a test statistic larger in magnitude than the observed test statistic if the null hypothesis is true and the probability model is accurate.

More generally, we can say that the p-value is the probability of getting a test statistic as extreme or more extreme than the observed test statistic in the direction of H_A if the null hypothesis is true and the probability model is accurate.

When working with proportions, we can also say that the p-value is the probability of getting a sample proportion as far from or farther from the null proportion in the direction of H_A if the null hypothesis is true and the normal model holds.

When the p-value is small, i.e. less than a previously set threshold, we say the results are **statistically significant**. This means the data provide such strong evidence against H_0 that we reject the null hypothesis in favor of the alternative hypothesis. The threshold is called the **significance level** and is represented by α (the Greek letter *alpha*). The significance level is typically set to $\alpha = 0.05$, but can vary depending on the field or the application.

STATISTICAL SIGNIFICANCE

If the p-value is less than the significance level α (usually 0.05), we say that the result is **statistically significant**. We reject H_0, and we have strong evidence favoring H_A.

If the p-value is greater than the significance level α, we say that the result is not statistically significant. We do not reject H_0, and we do not have strong evidence for H_A.

Recall that the null claim is the claim of no difference. If we reject H_0, we are asserting that there is a real difference. If we do not reject H_0, we are saying that the null claim is reasonable, but we are not saying that the null claim has been proven.

GUIDED PRACTICE 5.28

Because the p-value is 0.1754, which is larger than the significance level 0.05, we do not reject the null hypothesis. Explain what this means in the context of the problem using plain language.[15]

---

[15]The data do not provide evidence that the consultant's complication rate is significantly lower or higher than the US complication rate of 10%.

EXAMPLE 5.29

In the previous exercise, we did not reject H_0. This means that we did not disprove the null claim. Is this equivalent to proving the null claim is true?

No. We did not prove that the consultant's complication rate is *exactly* equal to 10%. Recall that the test of hypothesis starts by *assuming the null claim is true.* That is, the test proceeds as an argument by contradiction. *If the null claim is true*, there is a 0.1754 chance of seeing sample data as divergent from 10% as we saw in our sample. Because 0.1754 is large, it is within the realm of chance error, and we cannot say the null hypothesis is unreasonable.[16]

DOUBLE NEGATIVES CAN SOMETIMES BE USED IN STATISTICS

In many statistical explanations, we use double negatives. For instance, we might say that the null hypothesis is *not implausible* or we *failed to reject* the null hypothesis. Double negatives are used to communicate that while we are not rejecting a position, we are also not saying that we know it to be true.

EXAMPLE 5.30

Does the conclusion in Guided Practice 5.28 ensure that there is no real association between the surgical consultant's work and the risk of complications? Explain.

No. It is possible that the consultant's work is associated with a lower or higher risk of complications. If this was the case, the sample may have been too small to reliable detect this effect.

EXAMPLE 5.31

An experiment was conducted where study participants were randomly divided into two groups. Both were given the opportunity to purchase a DVD, but one half was reminded that the money, if not spent on the DVD, could be used for other purchases in the future, while the other half was not. The half that was reminded that the money could be used on other purchases was 20% less likely to continue with a DVD purchase. We determined that such a large difference would only occur about 1-in-150 times if the reminder actually had no influence on student decision-making. What is the p-value in this study? Was the result statistically significant?

The p-value was 0.006 (about 1/150). Since the p-value is less than 0.05, the data provide statistically significant evidence that US college students were actually influenced by the reminder.

WHAT'S SO SPECIAL ABOUT 0.05?

We often use a threshold of 0.05 to determine whether a result is statistically significant. But why 0.05? Maybe we should use a bigger number, or maybe a smaller number. If you're a little puzzled, that probably means you're reading with a critical eye – good job! We've made a video to help clarify *why 0.05*:

www.openintro.org/why05

Sometimes it's a good idea to deviate from the standard. We'll discuss when to choose a threshold different than 0.05 in Section 5.3.7.

Statistical inference is the practice of making decisions and conclusions from data in the context of uncertainty. Just as a confidence interval may occasionally fail to capture the true value of the parameter, a test of hypothesis may occasionally lead us to an incorrect conclusion. While a given data set may not always lead us to a correct conclusion, statistical inference gives us tools to control and evaluate how often these errors occur.

---

[16]The p-value is a conditional probability. It is P(getting data at least as divergent from the null value as we observed | H_0 is true). It is NOT P(H_0 is true | we got data this divergent from the null value).

5.3.4 Calculating the p-value by simulation (special topic)

When conditions for the applying a normal model are met, we use a normal model to find the p-value of a test of hypothesis. In the complication rate example, the distribution is not normal. It is, however, *binomial*, because we are interested in how many out of 142 patients will have complications.

We could calculate the p-value of this test using binomial probabilities. A more general approach, though, for calculating p-values when a normal model does not apply is to use what is known as **simulation**. While performing this procedure is outside of the scope of the course, we provide an example here in order to better understand the concept of a p-value.

We simulate 142 new patients to see what result might happen if the complication rate really is 0.10. To do this, we could use a deck of cards. Take one red card, nine black cards, and mix them up. If the cards are well-shuffled, drawing the top card is one way of simulating the chance a patient has a complication if the true rate is 0.10: if the card is red, we say the patient had a complication, and if it is black then we say they did not have a complication. If we repeat this process 142 times and compute the proportion of simulated patients with complications, $\hat{p}_{sim}$, then this simulated proportion is exactly a draw from the null distribution.

There were 12 simulated cases with a complication and 130 simulated cases without a complication: $\hat{p}_{sim} = 12/142 = 0.085$.

One simulation isn't enough to get a sense of the null distribution, so we repeated the simulation 10,000 times using a computer. Figure 5.9 shows the null distribution from these 10,000 simulations. The simulated proportions that are less than or equal to $\hat{p} = 0.063$ are shaded. There were 0.0877 simulated sample proportions with $\hat{p}_{sim} \leq 0.063$, which represents a fraction 0.0877 of our simulations:

$$\text{left tail} \ = \ \frac{\text{Number of observed simulations with } \hat{p}_{sim} \leq 0.063}{10000} = \frac{877}{10000} = 0.0877$$

However, this is not our p-value! Remember that we are conducting a two-sided test, so we should double the one-tail area to get the p-value:[17]

$$\text{p-value} = 2 \times \text{left tail} = 2 \times 0.0877 = 0.1754$$
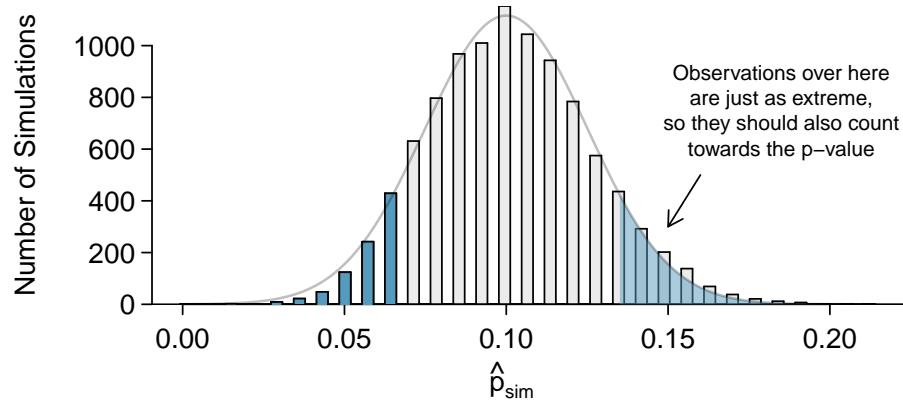
Figure 5.9: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail contains 8.77% of the simulations. For a two-sided test, we double the tail area to get the p-value. This doubling accounts for the observations we might have observed in the upper tail, which are also at least as extreme (relative to 0.10) as what we observed, $\hat{p} = 0.063$.

---

[17]This doubling approach is preferred even when the distribution isn't symmetric, as in this case.

5.3.5   Hypothesis testing: a five step process

Use a hypothesis test to *test H_0 versus H_A* at a particular *signficance level, α*.

(AP EXAM TIP) WHEN CARRYING OUT A HYPOTHESIS TEST PROCEDURE, FOLLOW THESE FIVE STEPS:

- **Identify**: Identify the hypotheses and the significance level.
- **Choose**: Choose the appropriate test procedure and identify it by name.
- **Check**: Check that the conditions for the test procedure are met.
- **Calculate**: Calculate the test statistic and the p-value.

$$\text{test statistic} = \frac{\text{point estimate}\ -\ \text{null value}}{SE\ \text{of estimate}}$$

- **Conclude**: Compare the p-value to the significance level to determine whether to reject H_0 or not reject H_0. Draw a conclusion in the context of H_A.

5.3.6   Decision errors

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism. The hallmarks of hypothesis testing are also found in the US court system.

EXAMPLE 5.32

A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?

The jury considers whether the evidence is so convincing (strong) that there is evidence beyond a reasonable doubt of the person's guilt. That is, the starting assumption (null hypothesis) is that the person is innocent until evidence is presented that convinces the jury that the person is guilty (alternative hypothesis). In statistics, our evidence comes in the form of data, and we use the significance level to decide what is beyond a reasonable doubt.

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Notice that a jury finds a defendant either guilty or not guilty. They either reject the null claim or they do not reject the null claim. They never prove the null claim, that is, they never find the defendant innocent. If a jury finds a defendant *not guilty*, this does not necessarily mean the jury is confident in the person's innocence. They are simply not convinced of the alternative that the person is guilty.

This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as truth.* Failing to find strong evidence for the alternative hypothesis is not equivalent to providing evidence that the null hypothesis is true.

Hypothesis tests are not flawless. Just think of the court system: innocent people are sometimes wrongly convicted and the guilty sometimes walk free. Similarly, data can point to the wrong conclusion. However, what distinguishes statistical hypothesis tests from a court system is that our framework allows us to quantify and control how often the data lead us to the incorrect conclusion.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios in a hypothesis test, which are summarized in Figure 5.10.

	Test conclusion	
	do not reject H_0	reject H_0 in favor of H_A
Truth H_0 true	correct conclusion	Type I Error
H_A true	Type II Error	correct conclusion

Figure 5.10: Four different scenarios for hypothesis tests.

TYPE I AND TYPE II ERRORS

A **Type I Error** is rejecting H_0 when H_0 is actually true. When we reject the null hypothesis, it is possible that we make a Type I Error.

A **Type II Error** is failing to reject H_0 when H_A is actually true. When we dod not reject the null hypothesis, it is possible that we make a Type II Error.

EXAMPLE 5.33

In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type I Error represent in this context? What does a Type II Error represent? Figure 5.10 may be useful.

———

If the court makes a Type I Error, this means the defendant is innocent (H_0 true) but wrongly convicted. A Type II Error means the court failed to reject H_0 (i.e. failed to convict the person) when they were in fact guilty (H_A true).

EXAMPLE 5.34

How could we reduce the Type I Error rate in US courts? What influence would this have on the Type II Error rate?

———

To lower the Type I Error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type II Errors.

GUIDED PRACTICE 5.35

How could we reduce the Type II Error rate in US courts? What influence would this have on the Type I Error rate?[18]

GUIDED PRACTICE 5.36

A group of women bring a class action lawsuit that claims discrimination in promotion rates. What would a Type I Error represent in this context?[19]

These examples provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

---

[18]To lower the Type II Error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type I Error rate.

[19]We must first identify which is the null hypothesis and which is the alternative. The alternative hypothesis is the one that bears the burden of proof, so the null hypothesis is that there was no discrimination and the alternative hypothesis is that there was discrimination. Making a Type I Error in this context would mean that in fact there was no discrimination, even though we concluded that women were discriminated against. Notice that this does *not* necessarily mean something was wrong with the data or that we made a computational mistake. Sometimes data simply point us to the wrong conclusion, which is why scientific studies are often repeated to check initial findings.

5.3.7   Choosing a significance level

If H_0 is true, what is the probability that we will incorrectly reject it? In hypothesis testing, we perform calculations under the premise that H_0 is true, and we reject H_0 if the p-value is smaller than the significance level α. That is, α *is* the probability of making a Type I Error. The choice of what to make α is not arbitrary. It depends on the gravity of the consequences of a Type I Error.

RELATIONSHIP BETWEEN TYPE I AND TYPE II ERRORS

The probability of a Type I Error is called α and corresponds to the significance level of a test. The probability of a Type II Error is called β. As we make α smaller, β typically gets larger, and vice versa.

EXAMPLE 5.37

If making a Type I Error is especially dangerous or especially costly, should we choose a smaller significance level or a higher significance level?

———————

Under this scenario, we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence before we are willing to reject the null hypothesis. Therefore, we want a smaller significance level, maybe $\alpha = 0.01$.

EXAMPLE 5.38

If making a Type II Error is especially dangerous or especially costly, should we choose a smaller significance level or a higher significance level?

———————

We should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

SIGNIFICANCE LEVELS SHOULD REFLECT CONSEQUENCES OF ERRORS

The significance level selected for a test should reflect the real-world consequences associated with making a Type I or Type II Error. If a Type I Error is very dangerous, make α smaller.

5.3.8   Statistical power of a hypothesis test

When the alternative hypothesis is true, the probability of <u>not</u> making a Type II Error is called **power**. It is common for researchers to perform a power analysis to ensure their study collects enough data to detect the effects they anticipate finding. As you might imagine, if the effect they care about is small or subtle, then if the effect is real, the researchers will need to collect a large sample size in order to have a good chance of detecting the effect. However, if they are interested in large effect, they need not collect as much data.

The Type II Error rate β and the magnitude of the error for a point estimate are controlled by the sample size. As the sample size n goes up, the Type II Error rate goes down, and power goes up. Real differences from the null value, even large ones, may be difficult to detect with small samples. However, if we take a very large sample, we might find a statistically significant difference but the size of the difference might be so small that it is of no practical value.

Section summary

- A **hypothesis test** is a statistical technique used to evaluate competing claims based on data.

- The competing claims are called **hypotheses** and are often about population parameters (e.g. μ and p); they are never about sample statistics.

  - The **null hypothesis** is abbreviated H_0. It represents a skeptical perspective or a perspective of no difference or *no change*.

  - The **alternative hypothesis** is abbreviated H_A. It represents a new perspective or a perspective of a real difference or change. Because the alternative hypothesis is the stronger claim, it bears the burden of proof.

- The **logic of a hypothesis test**: In a hypothesis test, we begin by *assuming that the null hypothesis is true.* Then, we calculate how unlikely it would be to get a sample value as extreme as we actually got in our sample, assuming that the null value is correct. If this likelihood is too small, it casts doubt on the null hypothesis and provides evidence for the alternative hypothesis.

- We set a **significance level**, denoted α, which represents the threshold below which we will reject the null hypothesis. The most common significance level is $\alpha = 0.05$. If we require more evidence to reject the null hypothesis, we use a smaller α.

- After verifying that the relevant **conditions are met**, we can calculate the test statistic. The **test statistic** tells us *how many* standard errors the point estimate (sample value) is from the null value (i.e. the value hypothesized for the parameter in the null hypothesis). When investigating a single mean or proportion or a difference of means or proportions, the test statistic is calculated as: $\frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$.

- After the test statistic, we calculate the p-value. We find and interpret the p-value according to the nature of the alternative hypothesis. The three possibilities are:

  H_A: parameter $>$ null value.   The p-value corresponds to the area in the *upper tail*.

  H_A: parameter $<$ null value.   The p-value corresponds to the area in the *lower tail*.

  H_A: parameter $\neq$ null value.   The p-value corresponds to the area in *both tails*.

  The **p-value** is the probability of getting a test statistic as extreme or more extreme than the observed test statistic in the direction of H_A if the null hypothesis is true and the probability model is accurate.

- The conclusion or decision of a hypothesis test is based on whether the p-value is smaller or larger than the preset significance level α.

  - When the p-value $< \alpha$, we say the results are **statistically significant** at the α level and we have evidence of a real difference or change. The observed difference is beyond what would have been expected from chance variation alone. This leads us to reject H_0 and gives us evidence for H_A.

  - When the p-value $> \alpha$, we say the results are not statistically significant at the α level and we do not have evidence of a real difference or change. The observed difference was within the realm of expected chance variation. This leads us to not reject H_0 and does not give us evidence for H_A.

- AP exam tip: A full hypothesis test includes the following steps.

  1. **Identify**: Identify the hypotheses and the significance level.
  2. **Choose**: Choose the appropriate test procedure and identify it by name.
  3. **Check**: Check that the conditions for the test procedure are met.
  4. **Calculate**: Calculate the test statistic and the p-value.

  $$\text{test statistic} = \frac{\text{point estimate} \ - \ \text{null value}}{SE \text{ of estimate}}$$

  5. **Conclude**: Compare the p-value to the significance level to determine whether to reject H_0 or not reject H_0. Draw a conclusion in the context of H_A.

- **Decision errors**. In a hypothesis test, there are two types of decision errors that could be made. These are called Type I and Type II Errors.

  - A **Type I Error** is rejecting H_0, when H_0 is actually true. We commit a Type I Error if we call a result significant when there is *no* real difference or effect. P(Type I error) = α.

  - A **Type II Error** is not rejecting H_0, when H_A is actually true. We commit a Type II Error if we call a result not significant when there *is* a real difference or effect. P(Type II error) = β.

  - The probability of a Type I Error (α) and a Type II Error (β) are *inversely related*. Decreasing α makes β larger; increasing α makes β smaller.

  - Once a decision is made, only one of the two types of errors is possible. If the test rejects H_0, for example, only a Type I Error is possible.

- The power of a test.

  - When a particular H_A is true, the probability of not making a Type II Error is called **power**. Power = $1 - \beta$.

  - The power of a test is the probability of detecting an effect of a particular size when it is present.

  - Increasing the significance level decreases the probability of a Type II Error and increases power. $\alpha \uparrow, \beta \downarrow, \text{power} \uparrow$.

  - For a fixed α, increasing the sample size n makes it easier to detect an effect and therefore decreases the probability of a Type II Error and increases power. $n \uparrow, \beta \downarrow, \text{power} \uparrow$.

Exercises

5.13 Identify hypotheses, Part I. Write the null and alternative hypotheses in words and then symbols for each of the following situations.

(a) A tutoring company would like to understand if most students tend to improve their grades (or not) after they use their services. They sample 200 of the students who used their service in the past year and ask them if their grades have improved or declined from the previous year.

(b) Employers at a firm are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non-business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity changed during March Madness.

5.14 Identify hypotheses, Part II. Write the null and alternative hypotheses in words and using symbols for each of the following situations.

(a) Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?

(b) The state of Wisconsin would like to understand the fraction of its adult residents that consumed alcohol in the last year, specifically if the rate is different from the national rate of 70%. To help them answer this question, they conduct a random sample of 852 residents and ask them about their alcohol consumption.

5.15 Online communication. A study suggests that 60% of college student spend 10 or more hours per week communicating with others online. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. You randomly sample 160 students from your dorm and find that 70% spent 10 or more hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \hat{p} < 0.6$$
$$H_A : \hat{p} > 0.7$$

5.16 Married at 25. A study suggests that the 25% of 25 year olds have gotten married. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. From a random sample of 25 year olds in census data with size 776, you find that 24% of them are married. A friend of yours offers to help you with setting up the hypothesis test and comes up with the following hypotheses. Indicate any errors you see.

$$H_0 : \hat{p} = 0.24$$
$$H_A : \hat{p} \neq 0.24$$

5.17 Cyberbullying rates. Teens were surveyed about cyberbullying, and 54% to 64% reported experiencing cyberbullying (95% confidence interval).[20] Answer the following questions based on this interval.

(a) A newspaper claims that a majority of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

(b) A researcher conjectured that 70% of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

(c) Without actually calculating the interval, determine if the claim of the researcher from part (b) would be supported based on a 90% confidence interval?

---

[20]Pew Research Center, A Majority of Teens Have Experienced Some Form of Cyberbullying. September 27, 2018.

5.18 Waiting at an ER, Part II. Exercise 5.11 provides a 95% confidence interval for the mean waiting time at an emergency room (ER) of (128 minutes, 147 minutes). Answer the following questions based on this interval.

(a) A local newspaper claims that the average waiting time at this ER exceeds 3 hours. Is this claim supported by the confidence interval? Explain your reasoning.

(b) The Dean of Medicine at this hospital claims the average wait time is 2.2 hours. Is this claim supported by the confidence interval? Explain your reasoning.

(c) Without actually calculating the interval, determine if the claim of the Dean from part (b) would be supported based on a 99% confidence interval?

5.19 Minimum wage, Part 1. Do a majority of US adults believe raising the minimum wage will help the economy, or is there a majority who do not believe this? A Rasmussen Reports survey of 1,000 US adults found that 42% believe it will help the economy.[21] Conduct an appropriate hypothesis test to help answer the research question.

5.20 Getting enough sleep. 400 students were randomly sampled from a large university, and 289 said they did not get enough sleep. Conduct a hypothesis test to check whether this represents a statistically significant difference from 50%, and use a significance level of 0.01.

5.21 Working backwards, Part I. You are given the following hypotheses:

$$H_0 : p = 0.3$$
$$H_A : p \neq 0.3$$

We know the sample size is 90. For what sample proportion would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

5.22 Working backwards, Part II. You are given the following hypotheses:

$$H_0 : p = 0.9$$
$$H_A : p \neq 0.9$$

We know that the sample size is 1,429. For what sample proportion would the p-value be equal to 0.01? Assume that all conditions necessary for inference are satisfied.

5.23 Testing for Fibromyalgia. A patient named Diana was diagnosed with Fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn't initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better.

(a) Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.

(b) What is a Type 1 Error in this context?

(c) What is a Type 2 Error in this context?

5.24 Which is higher? In each part below, there is a value of interest and two scenarios (I and II). For each part, report if the value of interest is larger under scenario I, scenario II, or whether the value is equal under the scenarios.

(a) The standard error of $\hat{p}$ when (I) $n = 125$ or (II) $n = 500$.

(b) The margin of error of a confidence interval when the confidence level is (I) 90% or (II) 80%.

(c) The p-value for a Z-statistic of 2.5 calculated based on a (I) sample with $n = 500$ or based on a (II) sample with $n = 1000$.

(d) The probability of making a Type 2 Error when the alternative hypothesis is true and the significance level is (I) 0.05 or (II) 0.10.

---

[21]Rasmussen Reports survey, Most Favor Minimum Wage of $10.50 Or Higher, April 16, 2019.

5.4 Does it make sense?

It is the responsibility of the data scientist to know when the use of these inference procedures is appropriate and to correctly interpret the results. In this section, we look at considerations around the misuse or misinterpretation of these procedures.

Learning objectives

1. Understand the two general conditions for when the confidence interval and hypothesis testing procedures apply. Explain why these conditions are necessary.

2. Distinguish between statistically significant and practically significant. What role does sample size play here?

3. Recognize that not all statistically significant results correspond to real differences, due to Type I Errors. What role does the significance level α play here?

5.4.1 When to retreat

Statistical tools rely on conditions. When the conditions are not met, these tools are unreliable and drawing conclusions from them is treacherous. The conditions for these tools typically come in two forms.

- **The individual observations must be independent.** A random sample from less than 10% of the population ensures the observations are independent. In experiments, we generally require that subjects are randomized into groups. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.

- **Other conditions focus on sample size and skew.** For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then a normal model for the sample mean will fail.

Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a data scientist.[22] The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Finally, we caution that there may be no inference tools helpful when considering data that include unknown biases, such as convenience samples. For this reason, there are books, courses, and researchers devoted to the techniques of sampling and experimental design. See Sections 1.3-1.5 for basic principles of data collection.

---

[22]If you work at a university, then there may be campus consulting services to assist you. Alternatively, there are many private consulting firms that are also available for hire.

5.4.2   Statistical significance versus practical significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected. While we still say that difference is **statistically significant**, it might not be **practically significant**.

Statistically significant differences are sometimes so minor that they are not practically relevant. This is especially important to research: if we conduct a study, we want to focus on finding a meaningful result. We don't want to spend lots of money finding results that hold no practical value.

The role of a data scientist in conducting a study often includes planning the size of the study. The data scientist might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain some reasonable estimate for the standard deviation. With these important pieces of information, she would choose a sufficiently large sample size so that the power for the meaningful difference is perhaps 80% or 90%. While larger sample sizes may still be used, she might advise against using them in some cases, especially in sensitive areas of research.

5.4.3   Statistical significance versus a real difference

When a result is statistically significant at the $\alpha = 0.05$ level, we have evidence that the result is real. However, when there is no difference or effect, we can expect that 5% of the time the test conclusion will lead to a Type I Error and incorrectly reject the null hypothesis. Therefore we must beware of what is called p-hacking, in which researchers may test many, many hypotheses and then publish the ones that come out statistically significant. As we noted, we can expect 5% of the results to be significant when the null hypothesis is true and there really is no difference or effect.[23]

---

[23] The problem is even greater than p-hacking. In what has been called the "reproducibility crisis", researchers have failed to reproduce a large proportion of results that were found significant and were published in scientific journals. This problem highlights the importance of research that reproduces earlier work rather than taking the word of a single study.

Also keep in mind that the probability that a difference will be found to be significant given that there is no real difference is not the same as the probability that a difference is not real, given that it was found significant. Depending upon the veracity of the hypotheses tested, the latter can be upwards of 80%, leading some to assert that "most published research is false".

https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab

Section summary

The inference procedures in this book require *two conditions* to be met.

- The first is that some type of **random sampling** or **random assignment** must be involved. If this is not the case, the point statistic may be biased and may not follow the intended distribution. Moreover, without a random sample or random assignment, there is no way to accurately measure the standard error. (When sampling without replacement, the sample size should be less than 10% of the population size in order for the standard error formula to apply. In sample surveys, this condition is generally met.)

- The second condition focuses on **sample size** and **skew** to determine whether the point estimate follows the intended distribution.

Understanding what the term **statistically significant** does and does not mean.

- *A small percent of the time (α), a significant result will not be a real result.* If many tests are run, a small percent of them will produce significant results due to chance alone.[24]

- *With a very large sample, a significant result may point to a result that is real but unimportant.* With a larger sample, the power of a test increases and it becomes easier to detect a small difference. If an extremely large sample is used, the result may be statistically significant, but not be *practically significant.* That is, the difference detected may be so small as to be unimportant or meaningless.

---

[24]Similarly, if many confidence intervals are constructed, a small percent (100 - C%) of them will fail to capture a true value due to chance alone. A value outside the confidence interval is not an *impossible* value.

Chapter highlights

Statistical inference is the practice of making decisions from data in the context of uncertainty. In this chapter, we introduced two frameworks for inference: **confidence intervals** and **hypothesis tests**.

- Confidence intervals are used for *estimating* unknown population parameters by providing an *interval of reasonable values* for the unknown parameter with a certain level of confidence.

- Hypothesis tests are used to assess how reasonable a *particular* value is for an unknown population parameter by providing *degrees of evidence* against that value.

- The results of confidence intervals and hypothesis tests are, generally speaking, *consistent*.[25] That is:

    - Values that fall *inside* a 95% confidence interval (implying they are reasonable) will *not be rejected* by a test at the 5% significance level (implying they are reasonable), and vice-versa.

    - Values that fall *outside* a 95% confidence interval (implying they are not reasonable) will *be rejected* by a test at the 5% significance level (implying they are not reasonable), and vice-versa.

    - When the confidence level and the significance level add up to 100%, the conclusions of the two procedures are consistent.

- Many values fall inside of a confidence interval and will not be rejected by a hypothesis test. "Not rejecting H_0" is NOT equivalent to *accepting H_0*. When we "do not reject H_0", we are asserting that the null value is *reasonable*, not that the parameter is exactly *equal to* the null value.

- For a 95% confidence interval, 95% is not the probability that the true value lies inside the confidence interval (it either does or it doesn't). Likewise, for a hypothesis test, α is not the probability that H_0 is true (it either is or it isn't). In both frameworks, the probability is about what would happen in a random sample, not about what is true of the population.

- The confidence interval procedures and hypothesis tests described in this book should not be applied unless particular conditions (described in more detail in the following chapters) are met. If these procedures are applied when the conditions are not met, the results may be unreliable and misleading.

While a given data set may not always lead us to a correct conclusion, statistical inference gives us tools to *control and evaluate how often errors occur*.

---

[25]In the context of proportions there will be a small range of cases where this is not true. This is because when working with proportions, the SE used for confidence intervals and the SE used for tests are slightly different, as we will see in the next chapter.

Chapter exercises

5.25 Relaxing after work. The General Social Survey asked the question: "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans.[26] A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92).

(a) Interpret this interval in context of the data.

(b) Suppose another set of researchers reported a confidence interval with a larger margin of error based on the same sample of 1,155 Americans. How does their confidence level compare to the confidence level of the interval stated above?

(c) Suppose next year a new survey asking the same question is conducted, and this time the sample size is 2,500. Assuming that the population characteristics, with respect to how much time people spend relaxing after work, have not changed much within a year. How will the margin of error of the 95% confidence interval constructed based on data from the new survey compare to the margin of error of the interval stated above?

5.26 Minimum wage, Part 2. In Exercise 5.19, we learned that a Rasmussen Reports survey of 1,000 US adults found that 42% believe raising the minimum wage will help the economy. Construct a 99% confidence interval for the true proportion of US adults who believe this.

5.27 Testing for food safety. A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.

(a) Write the hypotheses in words.

(b) What is a Type 1 Error in this context?

(c) What is a Type 2 Error in this context?

(d) Which error is more problematic for the restaurant owner? Why?

(e) Which error is more problematic for the diners? Why?

(f) As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.

5.28 True or false. Determine if the following statements are true or false, and explain your reasoning. If false, state how it could be corrected.

(a) If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

(b) Decreasing the significance level (α) will increase the probability of making a Type 1 Error.

(c) Suppose the null hypothesis is $p = 0.5$ and we fail to reject H_0. Under this scenario, the true population proportion is 0.5.

(d) With large sample sizes, even small differences between the null value and the observed point estimate, a difference often called the effect size, will be identified as statistically significant.

5.29 Unemployment and relationship problems. A USA Today/Gallup poll asked a group of unemployed and underemployed Americans if they have had major problems in their relationships with their spouse or another close family member as a result of not having a job (if unemployed) or not having a full-time job (if underemployed). 27% of the 1,145 unemployed respondents and 25% of the 675 underemployed respondents said they had major problems in relationships as a result of their employment status.

(a) What are the hypotheses for evaluating if the proportions of unemployed and underemployed people who had relationship problems were different?

(b) The p-value for this hypothesis test is approximately 0.35. Explain what this means in context of the hypothesis test and the data.

---

[26]National Opinion Research Center, General Social Survey, 2018.

5.30   Nearsighted. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted. Conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate?

5.31   Nutrition labels. The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a confidence interval for the number of calories per bag of 128.2 to 139.8 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips?

5.32   CLT for proportions. Define the term "sampling distribution" of the sample proportion, and describe how the shape, center, and spread of the sampling distribution change as the sample size increases when $p = 0.1$.

5.33   Practical vs. statistical significance. Determine whether the following statement is true or false, and explain your reasoning: "With large sample sizes, even small differences between the null value and the observed point estimate can be statistically significant."

5.34   Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is $n = 50$, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been $n = 500$. Will your p-value increase, decrease, or stay the same? Explain.

5.35   Gender pay gap in medicine. A study examined the average pay for men and women entering the workforce as doctors for 21 different positions.[27]

(a) If each gender was equally paid, then we would expect about half of those positions to have men paid more than women and women would be paid more than men in the other half of positions. Write appropriate hypotheses to test this scenario.

(b) Men were, on average, paid more in 19 of those 21 positions. Complete a hypothesis test using your hypotheses from part (a).

---

[27]Lo Sasso AT et al. "The $16,819 Pay Gap For Newly Trained Physicians: The Unexplained Trend Of Men Earning More Than Women". In: *Health Affairs* 30.2 (2011).

Chapter 6

Inference for categorical data

In this chapter, we apply the methods and ideas from Chapter 5 in several contexts for categorical data. We'll start by revisiting what we learned for a single proportion, where a normal distribution can be used to model the uncertainty in the sample proportion. Next, we apply these same ideas to analyze the difference of two proportions using a normal model. Later in the chapter we will encounter contingency tables, and we will use a different distribution, though the core ideas of hypothesis testing remain the same.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

6.1 Inference for a single proportion

In this section, we will apply the inferential procedures introduced in Chapter 5 to the context of a single proportion, and we will explore how to do sample size calculations for data collection purposes. We will answer questions such as the following:

- Do greater than half of adults in the U.S. oppose nuclear energy?

- What percent of adults in the U.S. approve of the way the Supreme Court is handling its job?

- What is the standard error is associated with this estimate?

- How do we construct a confidence interval for this value?

- What sample size is required to estimate this within a 3% margin of error using a 95% confidence level?

Learning objectives

1. State and verify whether or not the conditions for inference on a proportion using a normal distribution are met.

2. Recognize that the success-failure condition and the standard error calculation are different for the test and for the confidence interval and explain why this is the case.

3. Carry out a complete hypothesis test and confidence interval procedure for a single proportion.

4. Find the minimum sample size needed to estimate a proportion with C% confidence and a margin of error no greater than a certain value.

5. Recognize that margin of error calculations only measure sampling error, and that other types of errors may be present.

6.1.1   Distribution of a sample proportion (review)

The distribution of a sample proportion, such as the distribution of all possible values for the proportion of people who share a particular opinion in a poll, was introduced in Section 4.5. When the sampling distribution of a sample proportion, $\hat{p}$, is approximately normal, we can use confidence intervals and hypothesis tests based on a normal distribution. We call these Z-intervals and Z-tests for short. Here, we review the conditions necessary for a sample proportion to be modeled using a normal distribution.

CONDITIONS FOR THE SAMPLING DISTRIBUTION OF $\hat{P}$ BEING NEARLY NORMAL

The sampling distribution of a sample proportion, $\hat{p}$, based on a random sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and

2. $np \geq 10$ and $n(1 - p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

6.1.2   Checking conditions for inference using a normal distribution

We can use a normal model for inference for a proportion when the observations are independent and the sampling distribution of the sample proportion is nearly normal. We check that these assumptions are reasonable by verifying the following conditions.

Independent. Observations can be considered independent when the data are collected from a *random process*, such as tossing a coin, or from a *random sample*. Without a random sample or process, the standard error formula would not apply, and it is unclear to what population the inference would apply. When sampling without replacement from a finite population, the observations can be considered independent when sampling less than 10% of the population.[1]

Nearly normal sampling distribution. We saw in Section 4.5 that the sampling distribution of a sample proportion will be nearly normal when the success-failure condition is met, i.e. when the expected number of success and failures are both at least 10.

In our examples, we generally sample from large populations, such as the United States. In these cases, we do not explicitly verify that the sample size is less than 10% of the population size. However, in borderline cases, one should remember to check this condition as well to ensure that the standard error estimate is reasonable.

---

[1]When sampling without replacement and sampling greater than 10% of the population, a modified standard error formula should be used.

6.1.3 Confidence intervals for a proportion

The Gallup organization began measuring the public's view of the Supreme Court's job performance in 2000, and has measured it every year since then with the question: "Do you approve or disapprove of the way the Supreme Court is handling its job?". In 2018, the Gallup poll randomly sampled 1,033 adults in the U.S. and found that 53% of them approved.[2] We know that 53% is just a point estimate. What range of values are reasonable estimates for the percent of the population that approved of the job the Supreme Court is doing? We can use the confidence interval procedure introduced in the previous chapter to answer this question, but first we must clearly identify the parameter we're trying to estimate and be sure that a Z-interval will be appropriate. The following examples walk through the various steps for carrying out a confidence interval procedure using the Gallup poll data.

EXAMPLE 6.1

Identify the population of interest and the parameter of interest for the Gallup poll about the U.S. Supreme Court.

Gallup sampled from U.S. adults, therefore the population of interest, and the population to which we can make an inference, is U.S. adults. We know the percent of the sample that said they approve of the job the Supreme Court is doing. However, we do not know what percent of the population would approve. The parameter of interest, which is unknown, is the percent of *all* U.S. adults that approve of the job the Supreme Court is doing. This is the quantity that we seek to estimate with the confidence interval.

EXAMPLE 6.2

Can the sample proportion $\hat{p}$ be modeled using a normal distribution?

In order to construct a Z-interval, the sample statistic must be able to be modeled using a normal distribution. Gallup took a random sample, so the first condition (the independence condition) is satisfied. We must also test the second condition (the success-failure condition) to ensure that the sample size is large enough for the central limit theorem to apply. The success-failure condition is met when np and $n(1-p)$ are at least 10. Since p is always unknown when constructing a confidence interval for p, we use the sample proportion $\hat{p}$ to check this condition. Here we have:

$$n\hat{p} = 1033(0.53) = 547 \text{ ("successes")}$$
$$n(1 - \hat{p}) = 1033(1 - 0.53) = 486 \text{ ("failures")}$$

The second condition is satisfied since 547 and 486 are both at least 10. With the two conditions satisfied, we can model the sample proportion $\hat{p}$ using a normal model and we can construct a Z-interval.

[2]https://news.gallup.com/poll/237269/supreme-court-approval-highest-2009.aspx

EXAMPLE 6.3

Calculate the point estimate and the SE of the estimate.

The point estimate for the unknown parameter p (the proportion of all U.S. adults) who approve of the job the Supreme Court is doing) is the sample proportion. The point estimate here is $\hat{p} = 0.53$.

Because the point estimate is the sample proportion, the SE of the estimate is the SE of $\hat{p}$. In Section 4.5, we learned that the formula for the standard deviation of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

The proportion p is unknown, so we use the sample proportion $\hat{p}$ to find the SE of $\hat{p}$.

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Here $\hat{p} = 0.53$ and $n = 1,033$, so the SE of the sample proportion is:

$$SE = \sqrt{\frac{0.53(1-0.53)}{1033}} = 0.016$$

EXAMPLE 6.4

Construct a 90% confidence interval for p, the proportion of all U.S. adults that approve of the job the Supreme Court is doing.

Recall that the general form of a confidence interval is:

$$\text{point estimate } \pm \text{ critical value} \times SE \text{ of estimate}$$

We have already found the point estimate and the SE of the estimate. Because we previously verified that $\hat{p}$ can be modeled using a normal distribution, the critical value is a $z^\star$. The $z^\star$ value can be found in the t-table on page 524, using the bottom row (∞), where the column corresponds to the confidence level. Here the confidence level is 90%, so $z^\star$=1.65. We can now construct the 90% confidence interval as follows.

$$\text{point estimate} \pm z^\star \times SE \text{ of estimate}$$
$$0.53 \pm 1.65 \times 0.016$$
$$= (0.504, \ 0.556)$$

We are 90% confident that the true proportion of U.S. adults who approve of the job the Supreme Court is doing is between 0.504 and .556.

EXAMPLE 6.5

Based on the interval, is there evidence that more than half of U.S. adults approve of the job the Supreme Court is doing?

The 90% confidence interval (0.504, 0.556) provides an interval of reasonable values for the parameter. The value 0.50 is not in the interval, therefore can be considered unreasonable. Because the *entire* interval is above 0.50, we do have evidence, at the 90% confidence level, that more than half of U.S. adults (at the time of this poll) approve of the job the Supreme Court is doing.

EXAMPLE 6.6

Do we have evidence at the 95% confidence level that more than half of U.S. adults approve of the job the Supreme Court is doing?

———————

First, we observe that a 95% confidence interval will be *wider* than a 90% confidence interval. For a 95% Z-interval, $z^\star = 1.96$. The 95% confidence interval is:

$$0.53 \pm 1.96 \times 0.016$$
$$= (0.499, \ 0.561)$$

Now, we see that 0.50 is just barely inside the interval, making it within the range of reasonable values. Therefore, we do not have evidence, at the 95% confidence level, that more than half of U.S. adults (at the time of this poll) approve of the job the Supreme Court is doing.

Notice that we come to a different conclusion based on different confidence levels, which may feel a little jarring. However, this will happen with real data, and it highlights why it is important to be explicit in identifying the confidence level being used.

Having worked through this example, we now summarize the steps for constructing a confidence interval for a proportion using the five step framework introduce in Chapter 5.

CONSTRUCTING A CONFIDENCE INTERVAL FOR A PROPORTION

To carry out a complete confidence interval procedure to estimate a single proportion p,

Identify: Identify the parameter and the confidence level, C%.

> The parameter will be a population proportion, e.g. the proportion of all U.S. adults that approve of the job the Supreme Court is doing.

Choose: Choose the correct interval procedure and identify it by name.

> Here we choose the **1-proportion Z-interval**.

Check: Check conditions for the sampling distribution of $\hat{p}$ to be nearly normal.

> 1. Data come from a random sample or random process.
> 2. $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (Make sure to plug in numbers.)

Calculate: Calculate the confidence interval and record it in interval form.

> point estimate $\pm \ z^\star \times SE$ of estimate
>
> > point estimate: the sample proportion $\hat{p}$
> >
> > SE of estimate: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
> >
> > $z^\star$: use a t-table at row ∞ and confidence level C%
>
> (___, ___)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

> We are C% confident that the true *proportion* of [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value of interest.

EXAMPLE 6.7

A February 2018 Marist Poll reports: "Many Americans (68%) think there is intelligent life on other planets."[3] The results were based on a random sample of 1,033 adults in the U.S. Does this poll provide evidence at the 95% confidence level that greater than half of all U.S. adults think there is intelligent life on other planets? Carry out a confidence interval procedure to answer this question. Use the five step framework to organize your work.

Identify: First we identify the parameter of interest. Here the parameter is the true proportion of U.S. adults that think there is intelligent life on other planets. We will estimate this at the 95% confidence level.

Choose: Because the parameter to be estimated is a single proportion, we will use a 1-proportion Z-interval.

Check: We must check that a Z-interval is appropriate, meaning that the sample proportion can be modeled using a normal distribution. The problem states that the data come from a random sample. Also, we must check the success-failure condition. Here, we have that $1033(.68) \geq 10$ and $1033(1 - 0.68) \geq 10$. Both conditions are met so we can proceed with a 1-proportion Z-interval.

Calculate: We will calculate the interval:

$$\text{point estimate } \pm \ z^\star \times SE \text{ of estimate}$$

The point estimate is the sample proportion: $\hat{p} = 0.68$

The SE of the sample proportion is: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.68(1-0.68)}{1033}} = 0.015$.

$z^\star$ is found using the t-table at row ∞ and confidence level C%.
For a 95% confidence level, $z^\star = 1.96$.

The 95% confidence interval is given by:

$$0.68 \pm 1.96 \times \sqrt{\frac{0.68(1 - 0.68)}{1033}}$$
$$0.68 \pm 1.96 \times 0.015$$
$$= (0.651, \ 0.709)$$

Conclude: We are 95% confident that the true *proportion* of U.S. adults that think there is intelligent life on other planets is between 0.651 and 0.709. Because the entire interval is above 0.5 we have evidence that greater than half of all U.S. adults think there is intelligent life on other planets.

GUIDED PRACTICE 6.8

True or False: There is a 95% probability that between 65.1% and 70.9% of U.S. adults think that there is intelligent life on other planets.[4]

---

[3]This estimate of 68% in 2018 was up from an estimate of 52% in 2005. http://maristpoll.marist.edu/212-are-americans-poised-for-an-alien-invasion

[4]False. The true percent of U.S. adults that think there is intelligent life on other planets either falls in that interval or it doesn't. A correct interpretation of the confidence level would be that if we were to repeat this process over and over, about 95% of the 95% confidence intervals constructed would contain the true value.

6.1.4 Calculator: the 1-proportion Z-interval

A calculator can be helpful for evaluating the final interval in the Calculate step. However, it should not be used as a substitute for understanding.

TI-83/84: 1-PROPORTION Z-INTERVAL

Use STAT, TESTS, 1-PropZInt.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose A:1-PropZInt.
4. Let x be the *number* of yeses (must be an integer).
5. Let n be the sample size.
6. Let C-Level be the desired confidence level.
7. Choose Calculate and hit ENTER, which returns

   (__,__) the confidence interval
   p̂   the sample proportion
   n   the sample size

CASIO FX-9750GII: 1-PROPORTION Z-INTERVAL

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. Choose the INTR option (F4 button).
3. Choose the Z option (F1 button).
4. Choose the 1-P option (F3 button).
5. Specify the interval details:
   - Confidence level of interest for C-Level.
   - Enter the number of successes, x.
   - Enter the sample size, n.
6. Hit the EXE button, which returns

   Left, Right   ends of the confidence interval
   p̂   sample proportion
   n   sample size

GUIDED PRACTICE 6.9

Using a calculator, evaluate the confidence interval from Example 6.7. Recall that we wanted to find a 95% confidence interval for the proportion of U.S. adults who think there is intelligent life on other planets. The sample percent was 68% and the sample size was 1,033.[5]

---

[5]Navigate to the 1-proportion Z-interval on the calculator. To find x, the number of yes responses in the sample, we multiply the sample proportion by the sample size. Here $0.68 \times 1033 = 702.44$. We must round this to an integer, so we use x = 702. Also, n =1033 and C-Level = 0.95. The calculator output of (0.651, 0.708) matches our previously computed interval of (0.651, 0.709) with minor rounding difference.

6.1.5 Choosing a sample size when estimating a proportion

Planning a sample size before collecting data is important. If we collect too little data, the standard error of our point estimate may be so large that the estimate is not very useful. On the other hand, collecting data in some contexts is time-consuming and expensive, so we don't want to waste resources on collecting more data than we need.

When considering the sample size, we want to put an upper bound on the margin of error. Recall that the **margin of error** is measured as the distance between the point estimate and the lower or upper bound of a confidence interval.

MARGIN OF ERROR

The margin of error of a confidence interval is given by:

$$\text{critical value } \times SE \text{ of estimate}$$

The margin of error tells us with a given confidence level how far off we expect our point estimate to be from the true value.

EXAMPLE 6.10

Suppose we are conducting a university survey to determine whether students support a $200 per year increase in fees to pay for a new football stadium. Find the smallest sample size n so that the margin of error of the point estimate $\hat{p}$ will be no larger than 0.04 when using a 95% confidence level.

———————

Because we are working with proportions, the critical value is a $z^{\star}$ value. We want the margin of error to be less than or equal to 0.04, so we have:

$$z^{\star} \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04$$

There are two unknowns in the inequality: p and n. If we have an estimate of p, perhaps from a similar survey, we could use that value. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this *worst case estimate* of $p = 0.5$ if no other estimate is available.

$$1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.04$$

$$1.96^2 \times \frac{0.5(1-0.5)}{n} \leq 0.04^2$$

$$1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} \leq n$$

$$600.25 \leq n$$

$$n = 601$$

The sample size must be an integer and we round up because n must be greater than or equal to 600.25. We need at least 601 participants to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

No estimate of the true proportion is required in sample size computations for a proportion. However, if we have a reliable estimate of the proportion, we should use it in place of the worst case estimate of 0.5.

EXAMPLE 6.11

A recent estimate of Congress' approval rating was 17%.[6] If another poll were taken, what minimum sample size does this estimate suggest should be used to have a margin of error no greater than 0.04 with 95% confidence?

We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \le 0.04$$
$$n \ge 338.8$$
$$n = 339$$

If the true proportion is 0.17, then 339 is the minimum sample size that will ensure a margin of error no greater than 0.04 with 95% confidence.

IDENTIFY A SAMPLE SIZE FOR A PARTICULAR MARGIN OF ERROR

When estimating a single proportion, we find the minimum sample size n needed to achieve a margin of error no greater than m with a specified confidence level as follows:

$$z^\star \times \sqrt{\frac{p(1-p)}{n}} \le m$$

where $z^\star$ depends on the confidence level. If no reliable estimate of p exists, use $p = 0.5$.

GUIDED PRACTICE 6.12

All other things being equal, what would we have to do to the sample size in order to halve the margin of error (decrease it by a factor of 2)?[7]

GUIDED PRACTICE 6.13

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate the proportion of these tires that will be rejected through quality control. The quality control team has previously found that about 6.2% of tires fail inspection.

(a) How many tires should the manager examine to estimate the failure rate of the new tire model to within 2% with a 90% confidence level?[8]

(b) What if the estimate of p is 1.7% rather than 6.2%?[9]

---

[6]news.gallup.com/poll/237176/snapshot-congressional-job-approval-july.aspx

[7]To decrease the error, we would need to increase the sample size. We note that $\sqrt{n}$ is in the denominator of the SE formula, so we would have to *quadruple* the sample size in order to decrease the SE by a factor of 2. The margin of error as well as the width of the confidence interval is proportional to $\frac{1}{\sqrt{n}}$.

[9]The $z^\star$ corresponding to a 90% confidence level is 1.645. Since we have an estimate for p of 6.2%, we use it. So we have: $1.645 \times \sqrt{\frac{0.062(1-0.062)}{n}} \le 0.02$. Rearranging for n gives: $n \ge 393.4$, so she should use $n = 394$.

[9]Substituting 0.017 for p gives an n of 114. We can note that in this case $n \times p = 114 \times 0.017 = 1.9 < 10$. Since the success-failure condition is not met, the use of $z^\star = 1.645$ based on a normal model is not appropriate. We would need additional methods than what we've covered so far to get a good estimate for the minimum sample size in this scenario.

6.1.6   Hypothesis testing for a proportion

While a confidence interval provides a range of reasonable values for an unknown parameter, a hypothesis test evaluates a specific claim. In a hypothesis test, we set up competing hypotheses and find degrees of evidence against the null hypothesis.

EXAMPLE 6.14

Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. A newspaper collects a random sample of 500 likely voters in the district and estimates Toohey's support to be 52%.

(a) Identify the null and the alternative hypothesis. What value should we use as the null value, p_0?

(b) Can we model $\hat{p}$ using a normal model? Check the conditions.

(a) The alternative hypothesis, the one that bears the burden of proof, argues that Toohey has more than 50% support. Therefore, H_A will be one-sided and the null value will be $p_0 = 0.5$. So we have H_0: $p = 0.5$ and H_A: $p > 0.5$. Note that the hypotheses are about a population parameter. The hypotheses are never about the sample.

(b) First, we observe that the problem states that a random sample was chosen. Next, we check the success-failure condition. Because we assume that $p = p_0$ for the calculations of the hypothesis test, we use the hypothesized value p_0 rather than the sample value $\hat{p}$ when verifying the success-failure condition.

$$np_0 \geq 10 \quad \rightarrow \quad 500(0.5) \geq 10$$
$$n(1 - p_0) \geq 10 \quad \rightarrow \quad 500(1 - 0.5) \geq 10$$

The conditions for a normal model are met.

In Chapter 5, we saw that the general form of the test statistic for a hypothesis test takes the following form:

$$\text{test statistic} = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

When the conditions for a normal model are met:

- We use Z as the test statistic and call the test a Z-test.

- The point estimate is the sample proportion $\hat{p}$ (just like for a confidence interval).

- Since we compute the test statistic assuming the null hypothesis (that $p = p_0$) is true, we compute the standard error of the sample proportion using the null value p_0.

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

CONFIDENCE INTERVALS VERSUS HYPOTHESIS TESTS FOR A SINGLE PROPORTION

1-proportion Z-interval

$$\text{Check:} \quad n\hat{p} \geq 10 \text{ and } n(1 - \hat{p}) \geq 10 \qquad SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

1-proportion Z-test

$$\text{Check:} \quad np_0 \geq 10 \text{ and } n(1 - p_0) \geq 10 \qquad SE = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

EXAMPLE 6.15

(Continues previous example). Deborah Toohey's campaign manager claimed she has more than 50% support from the district's electorate. A newspaper poll finds that 52% of 500 likely voters who were sampled support Toohey. Does this provide convincing evidence for the claim by Toohey's manager at the 5% significance level?

————————

We will use a one-sided test with the following hypotheses:

H_0: $p = 0.5$. Toohey's support is 50%.

H_A: $p > 0.5$. Toohey's manager is correct, and her support is higher than 50%.

We will use a significance level of $\alpha = 0.05$ for the test. We can compute the standard error as

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.5(1 - 0.5)}{500}} = 0.022$$

The test statistic can be computed as:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}} = \frac{0.52 - 0.50}{0.022} = 0.89$$

Because the alternative hypothesis uses a greater than sign ($>$), this is an upper-tail test. We find the area under the standard normal curve to the *right* of $Z = 0.89$. A figure featuring the p-value is shown in Figure 6.1 as the shaded region.
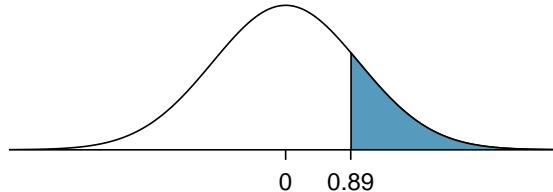
Figure 6.1: Sampling distribution of the sample proportion if the null hypothesis is true for Example 6.15. The p-value for the test is shaded.

Using a table or a calculator, we find the p-value is 0.19. This p-value of 0.19 is greater than $\alpha = 0.05$, so we do not reject H_0. That is, we do not have sufficient evidence to support Toohey's campaign manager's claims that she has more than 50% support within the district.

EXAMPLE 6.16

Based on the result above, do we have evidence that Toohey's support equals 50%?

————————

No. In a hypothesis test we look for degrees of evidence *against* the null hypothesis. We cannot ever prove the null hypothesis directly. The value 0.5 is reasonable, but many other values are reasonable as well. There are many values that would not get rejected by this test.

We now summarize the steps for carrying out a hypothesis test for a proportion using the five step framework introduced in the previous chapter.

HYPOTHESIS TESTING FOR A PROPORTION

To carry out a complete hypothesis test to test the claim that a single proportion p is equal to a null value p_0,

Identify: Identify the hypotheses and the significance level, α.

H_0: $p = p_0$

H_A: $p \neq p_0$; H_A: $p > p_0$;   or H_A: $p < p_0$

Choose: Choose the correct test procedure and identify it by name.

Here we choose the **1-proportion Z-test**.

Check: Check conditions for the sampling distribution of $\hat{p}$ to be nearly normal.

1. Data come from a random sample.
2. $np_0 \geq 10$ and $n(1 - p_0) \geq 10$ (Make sure to plug in numbers.)

Calculate: Calculate the Z-statistic and p-value.

$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

point estimate: the sample proportion $\hat{p}$

SE of estimate: $\sqrt{\frac{p_0(1-p_0)}{n}}$

null value: p_0

p-value = (based on the Z-statistic and the direction of H_A)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that $[H_A$ in context$]$.

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that $[H_A$ in context$]$.

EXAMPLE 6.17

A Gallup poll conducted in March of 2016 found that 54% of respondents oppose nuclear energy. This was the first time since Gallup first asked the question in 1994 that a majority of respondents said they oppose nuclear energy.[10] The survey was based on telephone interviews from a random sample of 1,019 adults in the United States. Does this poll provide evidence that greater than half of U.S. adults oppose nuclear energy? Carry out an appropriate test at the 0.10 significance level. Use the five step framework to organize your work.

———————

Identify: We will test the following hypotheses at the $\alpha = 10\%$ significance level.
   H_0: $p = 0.5$
   H_A: $p > 0.5$ Greater than half of all U.S. adults oppose nuclear energy.

   Note: $p > 0.5$ is what we want to find evidence for; this bears the burden of proof, so this corresponds to H_A.

Choose: Because the hypotheses are about a single proportion, we choose the 1-proportion Z-test.

Check: We must verify that the sample proportion can be modeled using a normal distribution. The problem states that the data come from a random sample. Also, $1019(0.5) \geq 10$ and $1019(1-0.5) \geq 10$ so both conditions are met. (Remember to use the hypothesized proportion, not the sample proportion, when checking the conditions for this test.)

Calculate: We will calculate the Z-statistic and the p-value.

$$Z = \frac{\text{point estimate} \ - \ \text{null value}}{SE \text{ of estimate}}$$

The point estimate is the sample proportion: $\hat{p} = 0.54$.

The value hypothesized for the parameter in H_0 is the null value: $p_0 = 0.5$

The SE of the sample proportion, assuming H_0 is true, is: $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1019}}$

$$Z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1019}}} = 2.5$$

Because H_A uses a greater than sign ($>$), meaning that it is an upper-tail test, the p-value is the area to the *right* of $Z = 2.5$ under the standard normal curve. This area can be found using a normal table or a calculator. The area or p-value = 0.006.

Conclude: The p-value of 0.006 is < 0.10, so we reject H_0; there is sufficient evidence that greater than half of U.S. adults oppose nuclear energy (as of March 2016).

GUIDED PRACTICE 6.18

In context, interpret the p-value of 0.006 from the previous example.[11]

———————

[10]www.gallup.com/poll/182180/support-nuclear-energy.aspx
[11]Assuming the normal model is accurate, there is a 0.006 probability of getting a test statistic greater than 2.5 *if H_0 were true*, that is, if the true proportion of U.S. adults that oppose nuclear energy really is 0.5. Note: We start by assuming H_0 is true, that p really equals 0.5. Then, assuming this, we estimate the probability of getting a sample proportion of 0.54 or larger by finding the area under the standard normal curve to the right of 2.5. This probability is very small, which casts doubt on the null hypothesis and leads us to reject it.

6.1.7   Calculator: the 1-proportion Z-test

A calculator can be useful for evaluating the test statistic and computing the p-value.

TI-83/84: 1-PROPORTION Z-TEST

Use `STAT`, `TESTS`, `1-PropZTest`.

1. Choose `STAT`.
2. Right arrow to `TESTS`.
3. Down arrow and choose `5:1-PropZTest`.
4. Let `p₀` be the null or hypothesized value of p.
5. Let `x` be the *number* of yeses (must be an integer).
6. Let `n` be the sample size.
7. Choose $\neq$, $<$, or $>$ to correspond to H_A.
8. Choose `Calculate` and hit `ENTER`, which returns

   z   Z-statistic
   p   p-value
   p̂   the sample proportion
   n   the sample size

CASIO FX-9750GII: 1-PROPORTION Z-TEST

The steps closely match those of the 1-proportion confidence interval.

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).
2. Choose the `TEST` option (`F3` button).
3. Choose the `Z` option (`F1` button).
4. Choose the `1-P` option (`F3` button).
5. Specify the test details:

   - Specify the sidedness of the test using the `F1`, `F2`, and `F3` keys.
   - Enter the null value, `p0`.
   - Enter the number of successes, `x`.
   - Enter the sample size, `n`.

6. Hit the `EXE` button, which returns

   z   Z-statistic
   p   p-value
   p̂   the sample proportion
   n   the sample size

GUIDED PRACTICE 6.19

Ⓖ   Using a calculator, find the test statistic and p-value for the earlier Example 6.17. Recall that we were looking for evidence that more than half of U.S. adults oppose nuclear energy. The sample percent was 54%, and the sample size was 1019.[12]

---

[12]Navigate to the 1-proportion Z-test on the calculator. Let `p₀` $= 0.5$. To find `x`, do $0.54 \times 1019 = 550.26$. This needs to be an integer, so round to the closest integer. Here `x` $= 550$. Also, `n` $= 1019$. We are looking for evidence that greater than half oppose, so choose $>$ `p₀`. When we do `Calculate`, we get the test statistic: `Z = 2.64` and the p-value: `p = 0.006`.

Section summary

Most of the confidence interval procedures and hypothesis tests of this book involve: a **point estimate**, the **standard error** of the point estimate, and an assumption about the **shape of the sampling distribution** of the point estimate. In this section, we explore inference when the parameter of interest is a *proportion*.

- We use the sample proportion $\hat{p}$ as the *point estimate* for the unknown population proportion p. The sampling distribution of $\hat{p}$ is approximately normal when the success-failure condition is met and the observations are independent. The observations can generally be considered independent when the data is collected from a random sample or come from a stable, random process analogous to flipping a coin. When the sampling distribution of $\hat{p}$ is normal, the standardized test statistic also follows a **normal** distribution.

- When verifying the success-failure condition and calculating the SE,
  - use the *sample* proportion $\hat{p}$ for the confidence interval, but
  - use the *null/hypothesized* proportion p_0 for the hypothesis test.

- When there is one sample and the parameter of interest is a single proportion:
  - Estimate p at the C% confidence level using a **1-proportion Z-interval**.
  - Test H_0: $p = p_0$ at the α significance level using a **1-proportion Z-test**.

- The first condition for the one proportion Z-interval and Z-test is the same. The second one is different because of the use of the null proportion for the test.

  1. The data come from a random sample or random process.
  2. Interval: $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ (Make sure to plug in numbers
     Test: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$ for n and $\hat{p}$, or for n and p_0!)

- When the conditions are met, we calculate the confidence interval and the test statistic as follows.

  Confidence interval: point estimate $\pm\ z^{\star} \times SE$ of estimate

  Test statistic: $Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the sample proportion $\hat{p}$.

  The SE of estimate is the SE of the sample proportion.

  For an Interval, use: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$;   for a Test, use: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

- The **margin of error** (ME) for a one-sample confidence interval for a proportion is $z^{\star}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is proportional to $\frac{1}{\sqrt{n}}$.

- To find the **minimum sample size** needed to estimate a proportion with a given confidence level and a given margin of error, m, set up an inequality of the form:

$$z^{\star}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < m$$

  $z^{\star}$ depends on the desired confidence level. Unless a particular proportion is given in the problem, use $\hat{p} = 0.5$. We solve for the sample size n. The final answer should be an *integer*, since n refers to a number of people or things.

Exercises

6.1 Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

(b) The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

(c) A random sample of 125 college students where 12% are vegetarians would be considered unusual.

(d) A random sample of 250 college students where 12% are vegetarians would be considered unusual.

(e) The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

6.2 Young Americans, Part I. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.[13]

(a) The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.

(b) The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

(c) A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

(d) A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

6.3 Orange tabbies. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

6.4 Young Americans, Part II. About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning.[14]

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

(b) In order for the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

---

[13]A. Vaughn. "Poll finds young adults optimistic, but not about money". In: *Los Angeles Times* (2011).

[14]Demos.org. "The State of Young America: The Poll". In: (2011).

6.5 Gender equality. The General Social Survey asked a random sample of 1,390 Americans the following question: "On the whole, do you think it should or should not be the government's responsibility to promote equality between men and women?" 82% of the respondents said it "should be". At a 95% confidence level, this sample has 2% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[15]

(a) We are 95% confident that between 80% and 84% of Americans in this sample think it's the government's responsibility to promote equality between men and women.

(b) We are 95% confident that between 80% and 84% of all Americans think it's the government's responsibility to promote equality between men and women.

(c) If we considered many random samples of 1,390 Americans, and we calculated 95% confidence intervals for each, 95% of these intervals would include the true population proportion of Americans who think it's the government's responsibility to promote equality between men and women.

(d) In order to decrease the margin of error to 1%, we would need to quadruple (multiply by 4) the sample size.

(e) Based on this confidence interval, there is sufficient evidence to conclude that a majority of Americans think it's the government's responsibility to promote equality between men and women.

6.6 Elderly drivers. The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% confidence level.[16]

(a) Verify the margin of error reported by The Marist Poll.

(b) Based on a 95% confidence interval, does the poll provide convincing evidence that *more than* 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?

6.7 Fireworks on July 4th. A local news outlet reported that 56% of 600 randomly sampled Kansas residents planned to set off fireworks on July 4th. Determine the margin of error for the 56% point estimate using a 95% confidence level.[17]

6.8 Life rating in Greece. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered "suffering".[18]

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions required for constructing a confidence interval based on these data are met.

(c) Construct a 95% confidence interval for the proportion of Greeks who are "suffering".

(d) Without doing any calculations, describe what would happen to the confidence interval if we decided to use a higher confidence level.

(e) Without doing any calculations, describe what would happen to the confidence interval if we used a larger sample.

6.9 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[19]

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% confidence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

(c) What does "90% confidence" mean?

(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

[15]National Opinion Research Center, General Social Survey, 2018.
[16]Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.
[17]Survey USA, News Poll #19333, data collected on June 27, 2012.
[18]Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.
[19]studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.10   Legalization of marijuana, Part I. The General Social Survey asked 1,578 US residents: "Do you think the use of marijuana should be made legal, or not?" 61% of the respondents said it should be made legal.[20]

(a) Is 61% a sample statistic or a population parameter? Explain.

(b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

(c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

(d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

6.11   National Health Plan, Part I. A *Kaiser Family Foundation* poll for US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan". There were 347 Democrats, 298 Republicans, and 617 Independents surveyed.[21]

(a) A political pundit on TV claims that a majority of Independents support a National Health Plan. Do these data provide strong evidence to support this type of statement?

(b) Would you expect a confidence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

6.12   Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.[22]

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a confidence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

6.13   Taste test. 🔳 Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then filled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identified the soda.

(a) Do these data provide strong evidence that these people are any better or worse than random guessing at telling the difference between diet and regular soda?

(b) Interpret the p-value in this context.

6.14   Is college worth it? Part II. Exercise 6.12 presents the results of a poll where 48% of 331 Americans who decide to not go to college do so because they cannot afford it.

(a) Calculate a 90% confidence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

(b) Suppose we wanted the margin of error for the 90% confidence level to be about 1.5%. How large of a survey would you recommend?

6.15   National Health Plan, Part II. Exercise 6.11 presents the results of a poll evaluating support for a generic "National Health Plan" in the US in 2019, reporting that 55% of Independents are supportive. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.16   Legalize Marijuana, Part II. As discussed in Exercise 6.10, the General Social Survey reported a sample where about 61% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey?

---

[20]National Opinion Research Center, General Social Survey, 2018.

[21]Kaiser Family Foundation, The Public On Next Steps For The ACA And Proposals To Expand Coverage, data collected between Jan 9-14, 2019.

[22]Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.

6.2 Difference of two proportions

We often wish two compare to groups to each other. In this section, we will answer the following questions:

- How much more effective is a blood thinner than a placebo for those who undergo CPR for a heart attack?

- How different is the approval of the 2010 healthcare law under two different question phrasings?

- Does the use of fish oils reduce heart attacks better than a placebo?

Learning objectives

1. State and verify whether or not the conditions for inference on the difference of two proportions using a normal distribution are met.

2. Recognize that the standard error calculation is different for the test and for the interval, and explain why that is the case.

3. Know how to calculate the pooled proportion and when to use it.

4. Carry out a complete confidence interval procedure for the difference of two proportions.

5. Carry out a complete hypothesis test for the difference of two proportions.

6.2.1 Sampling distribution of the difference of two proportions

In this section we want to compare two proportions to each other. We can start by taking their difference. If the difference is positive it tells us that the first one is larger. If it is negative, it tells use that the second one is larger. If the difference is zero, it tells us that they are equal. When comparing two proportions, then, the quantity that we want to estimate is really the difference: $p_1 - p_2$. This tells us how far apart the two proportions are.

Before we find a test statistic and perform inference for the two proportion case, we must investigate the sampling distribution of $\hat{p}_1 - \hat{p}_2$, which will become our point estimate. We know that the sampling distribution should be centered on $p_1 - p_2$. The standard deviation of $\hat{p}_1 - \hat{p}_2$ can be computed as:

$$SD_{\hat{p}_1-\hat{p}_2} = \sqrt{(SD_{\hat{p}_1})^2 + (SD_{\hat{p}_2})^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Like with $\hat{p}$, the difference of two sample proportions $\hat{p}_1 - \hat{p}_2$ follows a normal distribution when certain conditions are met. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

6.2.2   Checking conditions for inference using a normal distribution

When comparing two proportions, we carry out inference on $p_1 - p_2$. The assumptions are that the observations are independent, both between groups and within groups and that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is nearly normal. We check whether these assumptions are reasonable by verifying the following conditions.

Independent. Observations between groups can be considered independent when the data are collected from two independent random samples or, in the context of experiments, from two randomly assigned treatments. Randomly assigning subjects to treatments is equivalent to randomly assigning treatments to subjects. When sampling without replacement from a finite population, observations can be considered independent when sampling less than 10% of the population.

Nearly normal sampling distribution. The sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be nearly normal when the success-failure condition is met for both groups. In the two sample case, Instead of checking two inequalities, there are four to check.

6.2.3   Confidence interval for the difference of two proportions

We consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital. These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours. The results are shown in Figure 6.2.

	Survived	Died	Total
Treatment	14	26	40
Control	11	39	50
Total	25	65	90

Figure 6.2: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

Here, the parameter of interest is a difference of population proportions, specifically, the difference in the proportion of similar patients that would survive for at least 24 hours if in the treatment group versus if in the control group. Let:

$$p_1 : \text{proportion that would survive in treatment group, and}$$
$$p_2 : \text{proportion that would survive in control group}$$

Then the parameter of interest is $p_1 - p_2$. In order to use a Z-interval to estimate this difference, we must see if the point estimate, $\hat{p}_1 - \hat{p}_2$, follows a normal distribution. Because the patients were randomly assigned to one of the two groups and one heart attack patient is unlikely to influence the next that was in the study, the observations are considered independent, both within the samples and between the samples. Next, the success-failure condition should be verified for each group. We use the sample proportions along with the sample sizes to check the condition.

$$n_1\hat{p}_1 \geq 10 \qquad n_1(1 - \hat{p}_1) \geq 10 \qquad n_2\hat{p}_2 \geq 10 \qquad n_2(1 - \hat{p}_2) \geq 10$$
$$40 \times \frac{14}{40} \geq 10 \qquad 40 \times (1 - \frac{14}{40}) \geq 10 \qquad 50 \times \frac{11}{50} \geq 10 \qquad 50 \times (1 - \frac{11}{50}) \geq 10$$

Because all conditions are met, the normal model can be used for the point estimate of the difference in survival rate.

The point estimate is:

$$\hat{p}_1 - \hat{p}_2 = \frac{14}{40} - \frac{11}{50} = 0.35 - 0.22 = 0.13$$

We compute the standard error for the difference of sample proportions in the same way that we compute the standard deviation for the difference of sample proportions – the only difference is that we use the sample proportions in place of the population proportions:

$$SE = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} = \sqrt{\frac{0.35(1 - 0.35)}{40} + \frac{0.22(1 - 0.22)}{50}} = 0.095$$

Let us estimate the true difference in survival rate with 90% confidence. For a 90% confidence level, we use $z^\star = 1.645$. The 90% confidence interval is calculated as:

$$\text{point estimate} \pm z^\star \times SE \text{ of estimate}$$
$$0.13 \pm 1.65 \times 0.095$$
$$(-0.027,\ 0.287)$$

We are 90% confident that the true difference in the survival rate (treatment − control) lies between -0.027 and 0.095. That is, we are 90% confident that the treatment of blood thinners changes survival rate for patients like those in the study by -2.7% to +28.7% percentage points. Because this interval contains both negative and positive values, we do not have enough information to say with confidence whether blood thinners harm or help heart attack patients who have been admitted after they have undergone CPR.

CONSTRUCTING A CONFIDENCE INTERVAL FOR THE DIFFERENCE OF TWO PROPORTIONS

To carry out a complete confidence interval procedure to estimate the difference of two proportions $p_1 - p_2$,

Identify: Identify the parameter and the confidence level, C%.

> The parameter will be a difference of proportions, e.g. the true difference in the proportion of 17 and 18 year olds with a summer job (proportion of 18 year olds − proportion of 17 year olds).

Choose: Identify the correct interval procedure and identify it by name.

> Here we choose the **2-proportion Z-interval**.

Check: Check conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be nearly normal.

> 1. Data come from 2 independent random samples or 2 randomly assigned treatments.
> 2. $n_1\hat{p}_1 \geq 10$, $n_1(1 - \hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1 - \hat{p}_2) \geq 10$

Calculate: Calculate the confidence interval and record it in interval form.

> point estimate $\pm\ z^\star \times SE$ of estimate
>
> > point estimate: the difference of sample proportions $\hat{p}_1 - \hat{p}_2$
> >
> > SE of estimate: $\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$
> >
> > $z^\star$: use a t-table at row ∞ and confidence level C%
>
> (___, ___)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

> We are C% confident that the true *difference in the proportion* of [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.

EXAMPLE 6.20

A remote control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. The quality control engineer collects a sample of gears, examining 1000 gears from each company and finds that 879 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, construct a 95% confidence interval for the difference in the proportion from each supplier that would pass inspection. Use the five step framework described above to organize your work.

Identify: First we identify the parameter of interest. Here the parameter we wish to estimate is the true difference in the proportion of gears from each supplier that would pass inspection, $p_1 - p_2$. We will take the difference as: current − prospective, so p_1 is the true proportion that would pass from the current supplier and p_2 is the true proportion that would pass from the prospective supplier. We will estimate the difference using a 95% confidence level.

Choose: Because the parameter to be estimated is a difference of proportions, we will use a 2-proportion Z-interval.

Check: The samples are independent, but not necessarily random, so to proceed we must assume the gears are all independent. For this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as to whether this assumption is appropriate. We also must verify the minimum sample size conditions:

$$1000 \times \frac{879}{1000} \geq 10 \qquad 1000 \times \frac{121}{1000} \geq 10 \qquad 1000 \times \frac{958}{1000} \geq 10 \qquad 1000 \times \frac{42}{1000} \geq 10$$

The success-failure condition is met for both samples.

Calculate: We will calculate the interval:

$$\text{point estimate} \ \pm \ z^\star \times SE \text{ of estimate}$$

The point estimate is the difference of sample proportions: $\hat{p}_1 - \hat{p}_2 = 0.879 - 0.958 = -0.079$.

The SE of the difference of sample proportions is:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.879(1-0.879)}{1000} + \frac{0.958(1-0.958)}{1000}} = 0.0121$$

So the 95% confidence interval is given by:

$$0.879 - 0.958 \pm 1.96 \times \sqrt{\frac{0.879(1-0.879)}{1000} + \frac{0.958(1-0.958)}{1000}}$$
$$-0.079 \pm 1.96 \times 0.0121$$
$$(-0.103, \ -0.055)$$

Conclude: We are 95% confident that the true difference (current − prospective) in the proportion that would pass inspection is between -0.103 and -0.055, meaning that we are 95% confident that the prospective supplier would have between a 5.5% and 10.3% *greater* rate of passing inspection. Because the entire interval is below zero, the data provide sufficient evidence that the prospective gears pass inspection more often than the current gears. The remote control car company should go with the new manufacturer.

6.2.4   Calculator: the 2-proportion Z-interval

As with the 1-proportion Z-interval, a calculator can be helpful for evaluating the final interval.

TI-83/84: 2-PROPORTION Z-INTERVAL

Use STAT, TESTS, 2-PropZInt.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose B:2-PropZInt.
4. Let x1 be the *number* of yeses (must be an integer) in sample 1 and let n1 be the size of sample 1.
5. Let x2 be the *number* of yeses (must be an integer) in sample 2 and let n2 be the size of sample 2.
6. Let C-Level be the desired confidence level.
7. Choose Calculate and hit ENTER, which returns:
   (__,__)   the confidence interval
   $\hat{p}_1$ sample 1 proportion n_1 size of sample 1
   $\hat{p}_2$ sample 2 proportion n_2 size of sample 2

CASIO FX-9750GII: 2-PROPORTION Z-INTERVAL

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. Choose the INTR option (F4 button).
3. Choose the Z option (F1 button).
4. Choose the 2-P option (F4 button).
5. Specify the interval details:

   • Confidence level of interest for C-Level.
   • Enter the number of successes for each group, x1 and x2.
   • Enter the sample size for each group, n1 and n2.

6. Hit the EXE button, which returns
   Left, Right   the ends of the confidence interval
   p̂1, p̂2   the sample proportions
   n1, n2   sample sizes

GUIDED PRACTICE 6.21

From Example 6.20, we have that a quality control engineer collects a sample of gears, examining 1000 gears from each company and finds that 879 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Use a calculator to find a 95% confidence interval for the difference (current − prospective) in the proportion that would pass inspection.[23]

---

[23]Navigate to the 2-proportion Z-interval on the calculator. Let x1 = 879, n1 = 1000, x2 = 958, and n2 = 1000. C-Level is .95. This should lead to an interval of (-0.1027, -0.0553), which matches what we found previously.

6.2.5   Hypothesis testing when H_0: $p_1 = p_2$

Here we use a new example to examine a special estimate of the standard error when the null hypothesis is that two population proportions equal each other, i.e. H_0: $p_1 = p_2$. We investigate whether the way a question is phrased can influence a person's response. Pew Research Center conducted a survey with the following question:[24]

> As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the original order given above, or they were reversed. Results are presented in Figure 6.3

	sample size	Approve law (%)	Disapprove law (%)	Other
"People who do not buy insurance will pay a penalty" is given first (original order)	771	47	49	3
"People who cannot afford it will receive financial help from the government" is given first (reversed order)	732	34	63	3

Figure 6.3: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

GUIDED PRACTICE 6.22

Is this study an experiment or an observational study?[25]

The approval percents of 47% and 34% seem far apart. However, could this difference be due to random chance? We will answer this question using a hypothesis test. To simplify things, let

p_1: the proportion of respondents that would approve of policy with the original statement ordering, and

p_2: the proportion of respondents that would approve of policy with the reversed statement ordering.

EXAMPLE 6.23

Set up hypotheses to test whether the two statement orders produce the same response.

The null claim is that the question order does not matter, that is, that the two proportions should be equal. The alternate claim, the one that bears the burden of proof, is that the question ordering does matter.

H_0: $p_1 = p_2$
H_A: $p_1 \neq p_2$

---

[24]www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate. sample sizes for each polling group are approximate.
[25]There is a random sample involved, but there are also two treatments. Half of the the respondents are given the original statement order and the other half, randomly, are given the reversed statement order. This is an experiment because there are randomly assigned treatments.

Now, we can note that:

$$p_1 = p_2 \text{ is equivalent to } p_1 - p_2 = 0, \text{ and}$$
$$p_1 \neq p_2 \text{ is equivalent to } p_1 - p_2 \neq 0.$$

We can now see that the hypotheses are really about a difference of proportions: $p_1 - p_2$. In the last section, we used a 2-proportion Z-interval to estimate the parameter $p_1 - p_2$; here, we will use a 2-proportion Z-test to test the null hypothesis that $p_1 - p_2 = 0$, i.e. that $p_1 = p_2$.

Recall that the test statistic Z has the form:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

The parameter of interest is $p_1 - p_2$, so the point estimate will be the observed difference of sample proportions: $\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13$.

The null value depends on the null hypothesis. The null hypothesis is that the approval rate would be the same for both statement orderings, i.e. that the difference is 0, therefore, the null value is 0. In this section we consider only the case where H_0: $p_1 = p_2$, so the null value for the difference will always be 0.

The SD of a difference of sample proportions has the form:

$$SD = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

However, in a hypothesis test, the distribution of the point estimate is always examined assuming the null hypothesis is true, i.e. in this case, $p_1 = p_2$. Both the success-failure check and the standard error formula should reflect this equality in the null hypothesis. We will use p_c to represent the common proportion that support healthcare law regardless of statement order:

$$SD = \sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}$$
$$= \sqrt{p_c(1 - p_c)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

We don't know the true proportion p_c, but we can obtain a good estimate of it, $\hat{p}_c$, by *pooling* the results of both samples. We find the total number of "yeses" or "successes" and divide that by the total number of cases. This is equivalent to taking a weighted average of $\hat{p}_1$ and $\hat{p}_2$. We call $\hat{p}_c$ the **pooled sample proportion**, and we use it to check the success-failure condition and to compute the standard error when the null hypothesis is that $p_1 = p_2$. Here:

$$\hat{p}_c = \frac{771(0.47) + 732(0.34)}{771 + 732} = 0.407$$

POOLED SAMPLE PROPORTION

When the null hypothesis is $p_1 = p_2$, it is useful to find the pooled sample proportion:

$$\hat{p}_c = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Here x_1 represents the number of successes in sample 1. If x_1 is not given, it can be computed as $n_1 \times \hat{p}_1$. Similarly, x_2 represents the number of successes in sample 2 and can be computed as $n_2 \times \hat{p}_2$.

> **USE THE POOLED SAMPLE PROPORTION WHEN H_0: $p_1 = p_2$**
>
> When the null hypothesis states that the proportions are equal, we use the pooled sample proportion $(\hat{p}_c)$ to check the success-failure condition and to estimate the standard error:
>
> $$SE = \sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

EXAMPLE 6.24

Verify that conditions for using the normal are met and find the SE of estimate for this hypothesis test. Recall that the pooled proportion $\hat{p}_c = 0.407$, $n_1 = 771$, and $n_2 = 732$.

The data do come from two randomly assigned treatments, where the treatments are the two different orderings of the question regarding healthcare. Also, the success-failure condition (minimums of 10) easily holds for each group.

$$771 \times 0.407 \geq 10 \qquad 771 \times (1 - 0.407) \geq 10 \qquad 732 \times 0.407 \geq 10 \qquad 732 \times (1 - 0.407) \geq 10$$

Here, we compute the SE for the difference of sample proportions as:

$$SE = \sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{0.407(1-0.407)}\sqrt{\frac{1}{771} + \frac{1}{732}} = 0.025$$

EXAMPLE 6.25

Complete the hypothesis test using a significance level of 0.01.

We have already set up the hypotheses and verified that the difference of proportions can be modeled using a normal distribution. We can now calculate the test statistic and p-value.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}} = \frac{(0.47 - 0.34) - 0}{0.025} = 5.2$$

This is a two-tailed test as H_A is that $p_1 \neq p_2$. We can find the area in one tail and double it. Here, the p-value ≈ 0. Because the p-value is smaller than $\alpha = 0.01$, we reject the null hypothesis and conclude that the order of the statements affects how likely a respondent is to support the 2010 healthcare law.

HYPOTHESIS TESTING FOR THE DIFFERENCE OF TWO PROPORTIONS

To carry out a complete hypothesis test to test the claim that two proportions p_1 and p_2 are equal to each other,

Identify: Identify the hypotheses and the significance level, α.

H_0: $p_1 = p_2$

H_A: $p_1 \neq p_2$; H_A: $p_1 > p_2$;   or H_A: $p_1 < p_2$

Choose: Identify the correct test procedure and identify it by name.

Here we choose the **2-proportion Z-test**.

Check: Check conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be nearly normal.

1. Data come from 2 independent random samples or from 2 randomly assigned treatments.

2. $n_1 \hat{p}_c \geq 10$, $n_1(1 - \hat{p}_c) \geq 10$, $n_2 \hat{p}_c \geq 10$, and $n_2(1 - \hat{p}_c) \geq 10$

Calculate: Calculate the Z-statistic and p-value.

$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

point estimate: the difference of sample proportions $\hat{p}_1 - \hat{p}_2$

SE of estimate: $\sqrt{\hat{p}_c(1 - \hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where $\hat{p}$ is the pooled proportion

null value: 0

p-value = (based on the Z-statistic and the direction of H_A)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].

EXAMPLE 6.26

A 5-year experiment was conducted to evaluate the effectiveness of fish oils on reducing heart attacks, where each subject was randomized into one of two treatment groups. We'll consider heart attack outcomes in these patients:

	heart_attack	no_event	Total
fish_oil	145	12788	12933
placebo	200	12738	12938

Carry out a complete hypothesis test at the 10% significance level to test whether the use of fish oils is effective in reducing heart attacks.

───────────

Identify: Define p_1 and p_2 as follows:
   p_1: the true proportion that would suffer a heart attack if given fish oil
   p_2: the true proportion that would suffer a heart attack if given placebo

   We will test the following hypotheses at the $\alpha = 0.10$ significance level.

   H_0: $p_1 = p_2$ Fish oil and placebo are equally effective.
   H_A: $p_1 < p_2$ Fish oil is effective in reducing heart attacks.

Choose: Because we are testing whether two proportions equal each other, we choose the 2-proportion Z-test.

Check: We must verify that the difference of sample proportions can be modeled using a normal distribution. First we note that there are two randomly assigned treatments. Second, we calculate the pooled proportion as follows:

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2} = \frac{145 + 200}{12933 + 12938} = 0.0133$$

We can now verify: $12933(0.0133) \geq 10$, $12933(1 - 0.0133) \geq 10$, $12938(0.0133) \geq 10$, and $12938(1 - 0.0133) \geq 10$, so both conditions are met.

Calculate: We will calculate the Z-statistic and the p-value.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

The point estimate is the difference of sample proportions: $\hat{p}_1 - \hat{p}_2 = 0.0112 - 0.0155 = -0.0043$.

The value hypothesized for the parameter in H_0 is the null value: null value $= 0$.

The pooled proportion, calculated above, is: $\hat{p}_c = 0.0133$.

The SE of the difference of sample proportions, assuming H_0 is true, is:
$$\sqrt{\hat{p}_c(1 - \hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{0.0133(1 - 0.0133)}\sqrt{\frac{1}{12933} + \frac{1}{12938}} = 0.00142.$$

$$Z = \frac{-0.0043 - 0}{0.00142} = -3.0$$

Because H_A uses a less than, meaning that it is a lower-tail test, the p-value is the area to the *left* of $Z = -3.0$ under the standard normal curve. This area can be found using a normal table or a calculator. The area or p-value $= 0.0013$.

Conclude: The p-value of 0.0013 is < 0.10, so we reject H_0; there is sufficient evidence that fish oil is effective in reducing heart attacks.

6.2.6 Calculator: the 2-proportion Z-test

▶ TI-83/84: 2-PROPORTION Z-TEST

Use STAT, TESTS, 2-PropZTest.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose 6:2-PropZTest.
4. Let x1 be the *number* of yeses (must be an integer) in sample 1 and let n1 be the size of sample 1.
5. Let x2 be the *number* of yeses (must be an integer) in sample 2 and let n2 be the size of sample 2.
6. Choose $\neq$, $<$, or $>$ to correspond to H_A.
7. Choose Calculate and hit ENTER, which returns:

z	Z-statistic	p	p-value
$\hat{p}_1$	sample 1 proportion	$\hat{p}$	pooled sample proportion
$\hat{p}_2$	sample 2 proportion		

▶ CASIO FX-9750GII: 2-PROPORTION Z-TEST

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. Choose the TEST option (F3 button).
3. Choose the Z option (F1 button).
4. Choose the 2-P option (F4 button).
5. Specify the test details:

   - Specify the sidedness of the test using the F1, F2, and F3 keys.
   - Enter the number of successes for each group, x1 and x2.
   - Enter the sample size for each group, n1 and n2.

6. Hit the EXE button, which returns

z	Z-statistic	$\hat{p}1$, $\hat{p}2$	sample proportions
p	p-value	$\hat{p}$	pooled proportion
		n1, n2	sample sizes

GUIDED PRACTICE 6.27

Use a calculator to find the test statistic, p-value, and pooled proportion for a test with:
H_A: p for fish oil $< p$ for placebo.[26]

	heart_attack	no_event	Total
fish_oil	145	12788	12933
placebo	200	12738	12938

---

[26]Correctly going through the calculator steps should lead to a solution with the test statistic z = -2.977 and the p-value p = 0.00145. These two values match our calculated values from the previous example to within rounding error. The pooled proportion is given as $\hat{p}$ = 0.0133. Note: values for x1 and x2 were given in the table. If, instead, proportions are given, find x1 and x2 by multiplying the proportions by the sample sizes and rounding the result to an *integer*.

Section summary

In the previous section, we looked at inference for a single proportion. In this section, we *compared* two groups to each other with respect to a proportion or a percent.

- We are interested in whether the true proportion of yeses is the same or different between two distinct groups. Call these proportions p_1 and p_2. The difference, $p_1 - p_2$ tells us whether p_1 is greater than, less than, or equal to p_2.

- When *comparing* two proportions to each other, the parameter of interest is the *difference of proportions*, $p_1 - p_2$, and we use the difference of sample proportions, $\hat{p}_1 - \hat{p}_2$, as the *point estimate*.

- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is nearly normal when the success-failure condition is met for *both* groups and when the data is collected using 2 independent random samples or 2 randomly assigned treatments. When the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is nearly normal, the standardized test statistic also follows a **normal** distribution.

- When the null hypothesis is that the two populations proportions are *equal* to each other, use the **pooled sample proportion** $\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$, i.e. the combined number of yeses over the combined sample sizes, when verifying the success-failure condition and when finding the SE. For the confidence interval, do not use the pooled sample proportion; use the separate values of $\hat{p}_1$ and $\hat{p}_2$.

- When there are two samples or treatments and the parameter of interest is a difference of proportions, e.g. the true difference in proportion of 17 and 18 year olds with a summer job (proportion of 18 year olds − proportion of 17 year olds):

  - Estimate $p_1 - p_2$ at the C% confidence level using a **2-proportion Z-interval**.
  - Test H_0: $p_1 - p_2 = 0$ (i.e. $p_1 = p_2$) at the α significance level using a **2-proportion Z-test**.

- Verify the conditions for using a normal model:

  1. Data come from 2 independent random samples or 2 randomly assigned treatments.
  2. CI: $n_1 \hat{p}_1 \geq 10$, $n_1(1 - \hat{p}_1) \geq 10$, $n_2 \hat{p}_2 \geq 10$, and $n_2(1 - \hat{p}_2) \geq 10$

     Test: $n_1 \hat{p}_c \geq 10$, $n_1(1 - \hat{p}_c) \geq 10$, $n_2 \hat{p}_c \geq 10$, and $n_2(1 - \hat{p}_c) \geq 10$

- When the conditions are met, we calculate the confidence interval and the test statistic using the same structure as in the previous section.

  Confidence interval: point estimate $\pm\ z^\star \times SE$ of estimate

  Test statistic: $Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the difference of sample proportions $\hat{p}_1 - \hat{p}_2$.

  The SE of estimate is the SE of a difference of sample proportions.

  For a CI, use: $SE = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$.

  For a Test, use: $SE = \sqrt{\hat{p}_c(1 - \hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Exercises

6.17 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

		Scenario		Total
		Provocative	Conservative	
	Yes	5	15	20
Intervene	No	15	10	25
	Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

6.18 Heart transplant success. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might benefit from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.[27]

	control	treatment
alive	4	24
dead	30	45

Suppose we are interested in estimating the difference in survival rate between the control and treatment groups using a confidence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

6.19 Gender and color preference. A study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.[28]

(a) We are 95% confident that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% confident that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% confidence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a significant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% confidence interval for ($p_{female} - p_{male}$) cannot be calculated with only the information given in this exercise.

---

[27]B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

[28]L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: *Personality and Individual Differences* 31.8 (2001), pp. 1375–1379.

6.20   The Daily Show. A Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. A 95% confidence interval for $(p_{\text{college grad}} - p_{\text{HS or less}})$, where p is the proportion of those who watch The Daily Show, is (0.07, 0.15). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[29]

(a) At the 5% significance level, the data provide convincing evidence of a difference between the proportions of college graduates and those with a high school degree or less who watch The Daily Show.

(b) We are 95% confident that 7% less to 15% more college graduates watch The Daily Show than those with a high school degree or less.

(c) 95% of random samples of 1,099 college graduates and 1,110 people with a high school degree or less will yield differences in sample proportions between 7% and 15%.

(d) A 90% confidence interval for $(p_{\text{college grad}} - p_{\text{HS or less}})$ would be wider.

(e) A 95% confidence interval for $(p_{\text{HS or less}} - p_{\text{college grad}})$ is (-0.15,-0.07).

6.21   National Health Plan, Part III. Exercise 6.11 presents the results of a poll evaluating support for a generically branded "National Health Plan" in the United States. 79% of 347 Democrats and 55% of 617 Independents support a National Health Plan.

(a) Calculate a 95% confidence interval for the difference between the proportion of Democrats and Independents who support a National Health Plan $(p_D - p_I)$, and interpret it in this context. We have already checked conditions for you.

(b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the National Health Plan than the Independent.

6.22   Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[30]

6.23   Offshore drilling, Part I. A survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[31]

(a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

6.24   Sleep deprivation, CA vs. OR, Part II. Exercise 6.22 provides data on sleep deprivation rates of Californians and Oregonians. The proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Conduct a hypothesis test to determine if these data provide strong evidence the rate of sleep deprivation is different for the two states. (Reminder: Check conditions)

(b) It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

---

[29]The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.
[30]CDC, Perceived Insufficient Rest or Sleep Among Adults — United States, 2008.
[31]Survey USA, Election Poll #16804, data collected July 8-11, 2010.

6.25 Offshore drilling, Part II. Results of a poll evaluating support for drilling for oil and natural gas off the coast of California were introduced in Exercise 6.23.

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample support drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who support off-shore drilling in California is different than that of non-college graduates.

6.26 Full body scan, Part I. A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1,137 adults nationwide, where one of the questions on the survey was "Some airports are now using 'full-body' digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?" Below is a summary of responses based on party affiliation.[32]

		Party Affiliation		
		Republican	Democrat	Independent
	Should	264	299	351
Answer	Should not	38	55	77
	Don't know/No answer	16	15	22
	Total	318	369	450

(a) Conduct an appropriate hypothesis test evaluating whether there is a difference in the proportion of Republicans and Democrats who think the full- body scans should be applied in airports. Assume that all relevant conditions are met.

(b) The conclusion of the test in part (a) may be incorrect, meaning a testing error was made. If an error was made, was it a Type 1 or a Type 2 Error? Explain.

6.27 Sleep deprived transportation workers. The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers. The results of the survey are shown below.[33]

			Transportation Professionals		
	Control	Pilots	Truck Drivers	Train Operators	Bus/Taxi/Limo Drivers
Less than 6 hours of sleep	35	19	35	29	21
6 to 8 hours of sleep	193	132	117	119	131
More than 8 hours	64	51	51	32	58
Total	292	202	203	180	210

Conduct a hypothesis test to evaluate if these data provide evidence of a difference between the proportions of truck drivers and non-transportation workers (the control group) who get less than 6 hours of sleep per day, i.e. are considered sleep deprived.

---

[32]S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: *CBS News* (2010).
[33]National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers' Sleep, 2012.

6.28 Prenatal vitamins and Autism. Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period).[34]

		Autism		
		Autism	Typical development	Total
Periconceptional	No vitamin	111	70	181
prenatal vitamin	Vitamin	143	159	302
	Total	254	229	483

(a) State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.

(b) Complete the hypothesis test and state an appropriate conclusion. (Reminder: Verify any necessary conditions for the test.)

(c) A New York Times article reporting on this study was titled "Prenatal Vitamins May Ward Off Autism". Do you find the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.[35]

6.29 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called *virologic failure*). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[36]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for difference in virologic failure rates between treatment groups.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: Verify any necessary conditions for the test.)

6.30 An apple a day keeps the doctor away. A physical education teacher at a high school wanting to increase awareness on issues of nutrition and health asked her students at the beginning of the semester whether they believed the expression "an apple a day keeps the doctor away", and 40% of the students responded yes. Throughout the semester she started each class with a brief discussion of a study highlighting positive effects of eating more fruits and vegetables. She conducted the same apple-a-day survey at the end of the semester, and this time 60% of the students responded yes. Can she used a two-proportion method from this section for this analysis? Explain your reasoning.

---

[34]R.J. Schmidt et al. "Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism". In: *Epidemiology* 22.4 (2011), p. 476.

[35]R.C. Rabin. "Patterns: Prenatal Vitamins May Ward Off Autism". In: *New York Times* (2011).

[36]S. Lockman et al. "Response to antiretroviral therapy after a single, peripartum dose of nevirapine". In: *Obstetrical & gynecological survey* 62.6 (2007), p. 361.

6.3 Testing for goodness of fit using chi-square

In this section, we develop a method for assessing a null model when the data take on more than two categories, such as yes/no/maybe instead of simply yes/no. This allows us to answer questions such as the following:

- Are juries representative of the population in terms of race/ethnicity, or is there a bias in jury selection?

- Is the color distribution of actual M&M's consistent with what was reported on the Mars website?

- Do people choose rock, paper, scissors with the same likelihood, or is one choice favored over another?

Learning objectives

1. Calculate the expected counts and degrees of freedom for a one-way table.

2. Calculate and interpret the test statistic χ^2.

3. State and verify whether or not the conditions for the chi-square goodness of fit are met.

4. Carry out a complete hypothesis test to evaluate if the distribution of a categorical variable follows a hypothesized distribution.

5. Understand how the degrees of freedom affect the shape of the chi-square curve.

6.3.1 Creating a test statistic for one-way tables

Data is collected from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Figure 6.4, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reflect the population of eligible jurors, i.e. registered voters.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Figure 6.4: Representation by race in a city's juries and population.

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

EXAMPLE 6.28

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

———————

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

GUIDED PRACTICE 6.29

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race? Answers can be found in Figure 6.5.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Figure 6.5: Actual and expected make-up of the jurors.

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

6.3.2   The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of point estimate}}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

In this example we have four categories: white, black, hispanic, and other. Because we have four values rather than just one or two, we need a new tool to analyze the data. Our strategy will be to find a test statistic that measures the overall deviation between the observed and the expected counts. We first find the difference between the observed and expected counts for the four groups:

	White	Black	Hispanic	Other
observed - expected	$205 - 198$	$26 - 19.25$	$25 - 33$	$19 - 24.75$

Next, we square the differences:

	White	Black	Hispanic	Other
(observed - expected)^2	$(205 - 198)^2$	$(26 - 19.25)^2$	$(25 - 33)^2$	$(19 - 24.75)^2$

We must standardize each term. To know whether the squared difference is large, we compare it to what was expected. If the expected count was 5, a squared difference of 25 is very large. However, if the expected count was 1,000, a squared difference of 25 is very small. We will divide each of the squared differences by the corresponding expected count.

	White	Black	Hispanic	Other
$\dfrac{\text{(observed - expected)}^2}{\text{expected}}$	$\dfrac{(205 - 198)^2}{198}$	$\dfrac{(26 - 19.25)^2}{19.25}$	$\dfrac{(25 - 33)^2}{33}$	$\dfrac{(19 - 24.75)^2}{24.75}$

Finally, to arrive at the overall measure of deviation between the observed counts and the expected counts, we add up the terms.

$$\chi^2 = \sum \frac{\text{(observed - expected)}^2}{\text{expected}}$$
$$= \frac{(205 - 198)^2}{198} + \frac{(26 - 19.25)^2}{19.25} + \frac{(25 - 33)^2}{33} + \frac{(19 - 24.75)^2}{24.75}$$

We can write an equation for χ^2 using the observed counts and expected counts:

$$\chi^2 = \frac{(\text{observed count}_1 - \text{expected count}_1)^2}{\text{expected count}_1} + \cdots + \frac{(\text{observed count}_4 - \text{expected count}_4)^2}{\text{expected count}_4}$$

The final number χ^2 summarizes how strongly the observed counts tend to deviate from the null counts.

In Section 6.3.4, we will see that if the null hypothesis is true, then χ^2 follows a new distribution called a *chi-square distribution*. Using this distribution, we will be able to obtain a p-value to evaluate whether there appears to be racial bias in the juries for the city we are considering.

6.3.3 The chi-square distribution and finding areas

The **chi-square distribution** is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall a normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

GUIDED PRACTICE 6.30

Figure 6.6 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[37]

Figure 6.6 and Guided Practice 6.30 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inflates.

---

[37](a) The center becomes larger. If we look carefully, we can see that the center of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly right skewed for $df = 2$, and then the distributions become more symmetric for the larger degrees of freedom $df = 4$ and $df = 9$. In fact, as the degrees of freedom increase, the χ^2 distribution approaches a normal distribution.
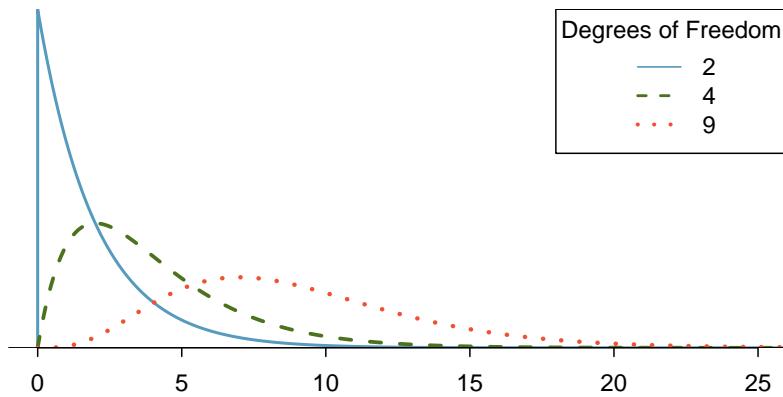
Figure 6.6: Three chi-square distributions with varying degrees of freedom.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the **chi-square table**, partially shown in Figure 6.7. A more complete table is presented in Appendix C.4 on page 526. This table is very similar to the t-table from Sections 7.1 and 7.3: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t-table is that the chi-square table only provides upper tail values.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	*6.25*	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Figure 6.7: A section of the chi-square table. A complete table is in Appendix C.4.

EXAMPLE 6.31

Figure 6.8(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Figure 6.7 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 6.8(a) has area 0.1.

EXAMPLE 6.32

We rarely observe the *exact* value in the table. For instance, Figure 6.8(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The lower bound for this upper tail is at 4.3, which does not fall in Figure 6.7. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 6.8(b) is between 0.1 and 0.2.

Using a calculator or statistical software allows us to get more precise areas under the chi-square curve than we can get from the table alone.
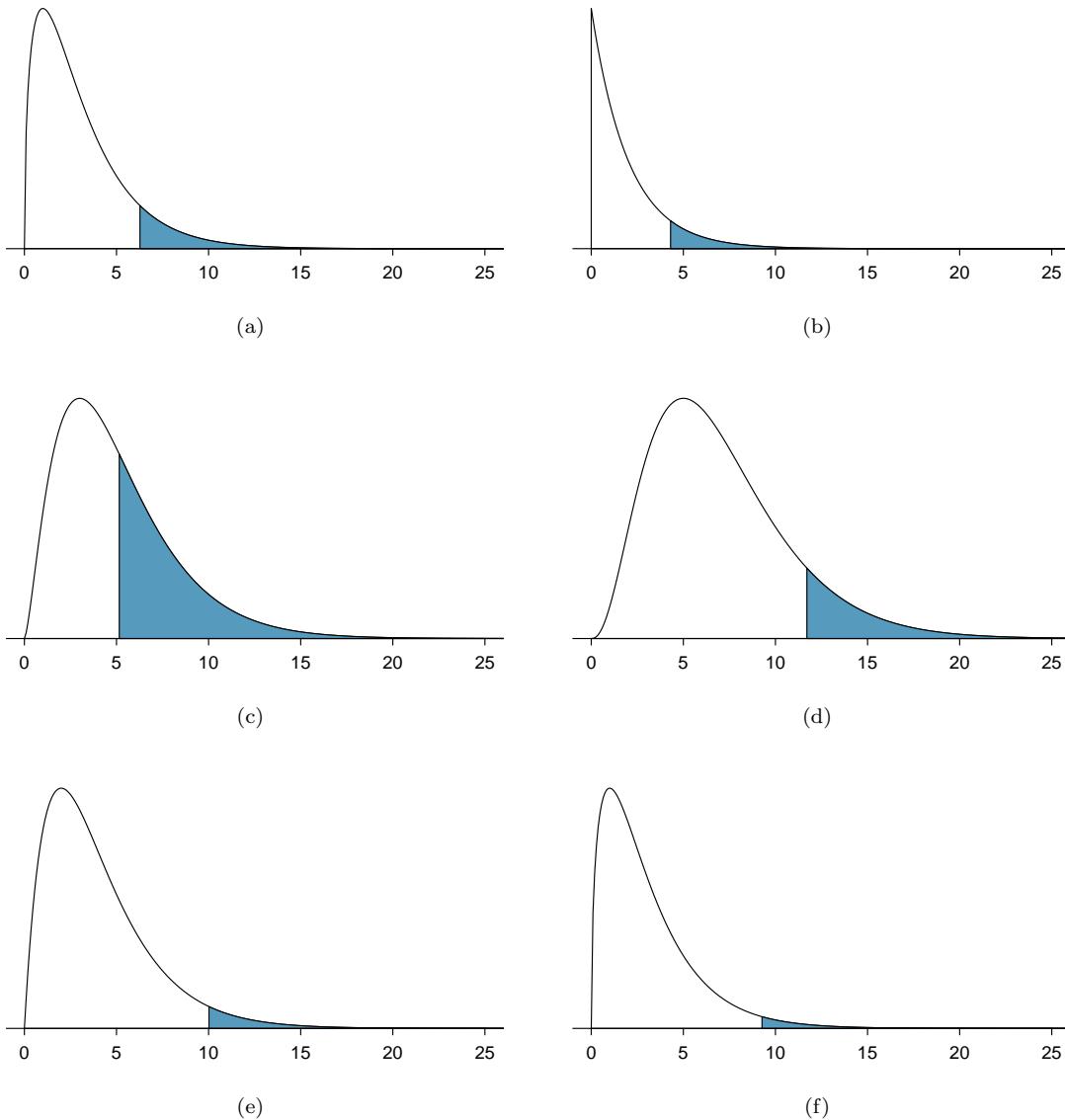
Figure 6.8: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

▶ TI-84: FINDING AN UPPER TAIL AREA UNDER THE CHI-SQUARE CURVE

Use the χ^2cdf command to find areas under the chi-square curve.

1. Hit `2ND VARS` (i.e. `DISTR`).
2. Choose `8:`χ^2`cdf`.
3. Enter the lower bound, which is generally the chi-square value.
4. Enter the upper bound. Use a large number, such as 1000.
5. Enter the degrees of freedom.
6. Choose `Paste` and hit `ENTER`.

TI-83: Do steps 1-2, then type the lower bound, upper bound, and degrees of freedom separated by commas. e.g. χ^2`cdf(5, 1000, 3)`, and hit `ENTER`.

▶ CASIO FX-9750GII: FINDING AN UPPER TAIL AREA UNDER THE CHI-SQ. CURVE

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).
2. Choose the `DIST` option (`F5` button).
3. Choose the `CHI` option (`F3` button).
4. Choose the `Ccd` option (`F2` button).
5. If necessary, select the `Var` option (`F2` button).
6. Enter the `Lower` bound (generally the chi-square value).
7. Enter the `Upper` bound (use a large number, such as 1000).
8. Enter the degrees of freedom, `df`.
9. Hit the `EXE` button.

GUIDED PRACTICE 6.33

(G) Figure 6.8(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area using a calculator.[38]

GUIDED PRACTICE 6.34

(G) Figure 6.8(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[39]

GUIDED PRACTICE 6.35

(G) Figure 6.8(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[40]

GUIDED PRACTICE 6.36

(G) Figure 6.8(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[41]

---

[38]Use a lower bound of 5.1, an upper bound of 1000, and $df = 5$. The upper tail area is 0.4038.
[39]The area is 0.1109.
[40]The area is 0.4043.
[41]The area is 0.0266.

6.3.4 Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (χ^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large χ^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($\chi^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then χ^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic χ^2 follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of bins or categories of the variable.

EXAMPLE 6.37

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for χ^2?

In the jurors example, there were $k = 4$ categories: white, black, Hispanic, and other. According to the rule above, the test statistic χ^2 should then follow a chi-square distribution with $k - 1 = 3$ degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely model χ^2 with a chi-square distribution. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can model the χ^2 test statistic, using a chi-square distribution.

EXAMPLE 6.38

If the null hypothesis is true, the test statistic $\chi^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value and state whether or not there is evidence of racial bias in the juror selection.

The chi-square distribution and p-value are shown in Figure 6.9. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using a calculator, we look at the chi-square curve with 3 degrees of freedom and find the area to the right of $\chi^2 = 5.89$. This area, which corresponds to the p-value, is equal to 0.117. This p-value is larger than the default significance level of 0.05, so we reject the null hypothesis. In other words, the data do not provide convincing evidence of racial bias in the juror selection.
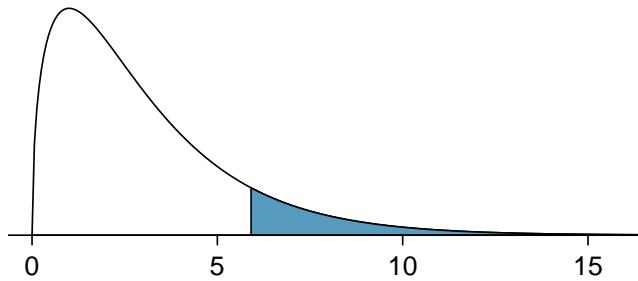
Figure 6.9: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

The test that we just carried out regarding jury selection is known as the **χ^2 goodness of fit test**. It is called "goodness of fit" because we test whether or not the proposed or expected distribution is a good fit for the observed data.

CHI-SQUARE GOODNESS OF FIT TEST FOR ONE-WAY TABLE

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts O_1, O_2, ..., O_k in k categories are unusually different from what might be expected under a null hypothesis. Calculate the *expected counts* that are based on the null hypothesis E_1, E_2, ..., E_k. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with $k - 1$ degrees of freedom:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of χ^2 would provide greater evidence against the null hypothesis.

CONDITIONS FOR THE CHI-SQUARE GOODNESS OF FIT TEST

The chi-square goodness of fit test requires two assumptions. The assumptions and the conditions that we check are listed below. If the conditions are not met, this test should not be used.

Independent. The observations can be considered independent if the data come from a random process. If randomly sampling from a finite population, the observations can be considered independent if sampling less than 10% of the population.

Sampling distribution is chi-square. In order for the χ^2-statistic to follow the chi-square distribution, each particular bin or category must have at least 5 expected cases under the assumption that the null hypothesis is true.

6.3.5 Evaluating goodness of fit for a distribution

GOODNESS OF FIT TEST FOR A ONE-WAY TABLE

When there is one sample and we are comparing the distribution of a categorical variable to a specified or population distribution, e.g. using sample values to determine if a machine is producing M&M's with the specified distribution of color,

Identify: Identify the hypotheses and the significance level, α.

H_0: The distribution of [...] matches the specified or population distribution.

H_A: The distribution of [...] doesn't match the specified or population distribution.

Choose: Choose the correct test procedure and identify it by name.

Here we choose the χ^2 **goodness of fit test**.

Check: Check that the test statistic follows a chi-square distribution.

1. Data come from a random sample or random process.
2. All expected counts are ≥ 5. (Make sure to calculate expected counts!)

Calculate: Calculate the χ^2-statistic, df, and p-value.

test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

$df = \#$ of categories $- 1$

p-value = (area to the *right* of χ^2-statistic with the appropriate df)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].

Have you ever wondered about the color distribution of M&M's®? If so, then you will be glad to know that Rick Wicklin, a statistician working at the statistical software company SAS, wondered about this too. But he did more than wonder; he decided to collect data to test whether the distribution of M&M colors was consistent with the stated distribution published on the Mars website in 2008. Starting at end of 2016, over the course of several weeks, he collected a sample of 712 candies, or about 1.5 pounds. We will investigate his results in the next example. You can read about his adventure in the Quartz article cited in the footnote below.[42]

---

[42]https://qz.com/918008/the-color-distribution-of-mms-as-determined-by-a-phd-in-statistics.

EXAMPLE 6.39

The stated color distribution of M&M's on the Mars website in 2008 is shown in the table below, along with the observed percentages from Rick Wicklin's sample of size 712. (See the paragraph before this example for more background.)

	Blue	Orange	Green	Yellow	Red	Brown
website percentages (2008):	24%	20%	16%	14%	13%	13%
observed percentages:	18.7%	18.7%	19.5%	14.5%	15.1%	13.5%

Is there evidence at the 5% significance level that the distribution of M&M's in 2016 were different from the stated distribution on the website in 2008? Use the five step framework to organize your work.

Identify: We will test the following hypotheses at the $\alpha = 0.05$ significance level.

H_0: The distribution of M&M colors is the same as the stated distribution in 2008.
H_A: The distribution of M&M colors is different than the stated distribution in 2008.

Choose: Because we have one variable (color), broken up into multiple categories, we choose the chi-square goodness of fit test.

Check: We must verify that the test statistic follows a chi-square distribution. Note that there is only one sample here. The website percentages are considered fixed – they are not the result of a sample and do not have sampling variability associated with them. To carry out the chi-square goodness of fit test, we will have to assume that Wicklin's sample can be considered a random sample of M&M's. Next, we need to find the expected counts. Here, $n = 712$. If H_0 is true, then we would expect 24% of the M&M's to be Blue, 20% to be Orange, etc. So the expected counts can be found as:

	Blue	Orange	Green	Yellow	Red	Brown
expected counts:	0.24(712) = 170.9	0.20(712) = 142.4	0.16(712) = 113.9	0.14(712) = 99.6	0.13(712) = 92.6	0.13(712) = 92.6

Calculate: We will calculate the chi-square statistic, degrees of freedom, and the p-value.
To calculate the chi-square statistic, we need the observed counts as well as the expected counts. To find the observed counts, we use the observed percentages. For example, 18.7% of $712 = 0.187(712) = 133$.

	Blue	Orange	Green	Yellow	Red	Brown
observed counts:	133	133	139	103	108	96
expected counts:	170.9	142.4	113.9	99.6	92.6	92.6

$$\chi^2 = \sum \frac{(\text{observed } - \text{ expected})^2}{\text{expected}}$$
$$= \frac{(133 - 170.9)^2}{170.9} + \frac{(133 - 142.4)^2}{142.4} + \cdots + \frac{(108 - 92.6)^2}{92.6} + \frac{(96 - 92.6)^2}{92.6}$$
$$= 8.41 + 0.62 + 5.53 + 0.12 + 2.56 + 0.12$$
$$= 17.36$$

Because there are six colors, the degrees of freedom is $6 - 1 = 5$.
In a chi-square test, the p-value is always the area to the *right* of the chi-square statistic. Here, the area to the right of 17.36 under the chi-square curve with 5 degrees of freedom is 0.004.

Conclude: The p-value of 0.004 is < 0.05, so we reject H_0; there is sufficient evidence that the distribution of M&M's does not match the stated distribution on the website in 2008.

EXAMPLE 6.40

For Wicklin's sample, which color showed the most prominent difference from the stated website distribution in 2008?

———————

We can compare the website percentages with the observed percentages. However, another approach is to look at the terms used when calculating the chi-square statistic. We note that the largest term, 8.41, corresponds to Blue. This means that the observed number for Blue was, relatively speaking, the farthest from the expected number among all of the colors. This is consistent with the observation that the largest difference in website percentage and observed percentage is for Blue (24% vs 18.7%). Wicklin observed far fewer Blue M&M's than would have been expected if the website percentages were still true.

6.3.6 Calculator: chi-square goodness of fit test

TI-84: CHI-SQUARE GOODNESS OF FIT TEST

Use `STAT`, `TESTS`, χ^2`GOF-Test`.

1. Enter the observed counts into list `L1` and the expected counts into list `L2`.

2. Choose `STAT`.

3. Right arrow to `TESTS`.

4. Down arrow and choose `D:`χ^2`GOF-Test`.

5. Leave `Observed: L1` and `Expected: L2`.

6. Enter the degrees of freedom after `df:`

7. Choose `Calculate` and hit `ENTER`, which returns:
   χ^2 chi-square test statistic
   `p`   p-value
   `df`   degrees of freedom

TI-83: Unfortunately the TI-83 does not have this test built in. To carry out the test manually, make list `L3 = (L1 - L2)`^{2} `/ L2` and do `1-Var-Stats` on `L3`. The sum of `L3` will correspond to the value of χ^2 for this test.

CASIO FX-9750GII: CHI-SQUARE GOODNESS OF FIT TEST

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).

2. Enter the observed counts into a list (e.g. `List 1`) and the expected counts into list (e.g. `List 2`).

3. Choose the `TEST` option (`F3` button).

4. Choose the `CHI` option (`F3` button).

5. Choose the `GOF` option (`F1` button).

6. Adjust the `Observed` and `Expected` lists to the corresponding list numbers from Step 2.

7. Enter the degrees of freedom, `df`.

8. Specify a list where the contributions to the test statistic will be reported using `CNTRB`. This list number should be different from the others.

9. Hit the `EXE` button, which returns
   χ^2 chi-square test statistic
   `p`   p-value
   `df`   degrees of freedom
   `CNTRB`   list showing the test statistic contributions

GUIDED PRACTICE 6.41

Use the table below and a calculator to find the χ^2-statistic and p-value for chi-square goodness of fit test.[43]

	Blue	Orange	Green	Yellow	Red	Brown
observed counts:	133	133	139	103	108	96
expected counts:	170.9	142.4	113.9	99.6	92.6	92.6

---

[43]Enter the observed counts into `L1` and the expected counts into `L2`. the `GOF` test. Make sure that `Observed:` is `L1` and `Expected:` is `L2`. Let `df:` be 5. You should find that $\chi^2 = 17.36$ and p-value $= 0.004$.

Section summary

The inferential procedures we saw in the first two sections of this chapter are based on the test statistic following a *normal distribution.* In this section, we introduce a new distribution called the chi-square distribution.

- While a normal distribution is defined by its mean and standard deviation, the chi-square distribution is defined by just one parameter called **degrees of freedom**.

- For a chi-square distribution, as the degrees of freedom increases:
  - the center increases.
  - the spread increases.
  - the shape becomes more symmetric and more normal.[44]

- When we want to see if a model is a good fit for observed data or if data is representative of a particular population, we can use a χ^2 **goodness of fit test**. This test is used when there is one variable with multiple categories (bins) that can be arranged in a **one-way table**.

- In a chi-square goodness of fit test, we calculate a χ^2**-statistic**, which is a measure of how far the observed values in the sample are from the expected values under the null hypothesis. $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$
  - Always use whole numbers (counts) for the observed values, not proportions or percents.
  - For each category, the expected counts can be found by multiplying the sample size by the expected proportion under the null hypothesis. Expected counts do *not* need to be integers.
  - For each category, find $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$, then add them all together to get the χ^2-statistic.

- When there is a random sample and all of the expected counts are at least 5, the χ^2-statistic follows a **chi-square distribution** with degrees of freedom equal to number of categories − 1.

- For a χ^2 test, the p-value corresponds to the probability that observed sample values would differ from the expected values by *more than* what we observed in this sample. The p-value, therefore, corresponds to the area *to the right* of the calculated χ^2-statistic (the area in the upper tail).

- A larger χ^2 represents greater deviation between the observed values and the expected values under the null hypothesis. For a fixed degrees of freedom, a larger χ^2 value leads to a smaller p-value, providing greater evidence against H_0.

- χ^2 **tests for a one-way table**. When there is one sample and we are comparing the distribution of a categorical variable to a specified or population distribution, e.g. using sample values to determine if a machine is producing M&M's with the specified distribution of color, the hypotheses can often be written as:

  H_0: The distribution of [...] matches the specified or population distribution.

  H_A: The distribution of [...] doesn't match the specified or population distribution.

  We test these hypotheses at the α significance level using a χ^2 **goodness of fit test**.

- The conditions for the χ^2 goodness of fit test are as follows:
  1. Data come from a random sample or random process.
  2. All expected counts are ≥ 5.

---

[44]Technically, however, it is always right skewed.

- We calculate the test statistic as follows:

$$\text{test statistic: } \chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}; \qquad df = \# \text{ of categories} - 1$$

- The p-value is the area to the *right* of the χ^2-statistic under the chi-square curve with the appropriate df.

Exercises

6.31 True or false, Part I. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

(c) The chi-square statistic is always positive.

(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

6.32 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found $\chi^2 = 10$ with $df = 5$ you would fail to reject H_0 at the 5% significance level.

(c) When finding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.33 Open source textbook. 📹 A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

(a) State the hypotheses for testing if the professor's predictions were inaccurate.

(b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

(c) This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

(d) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

(e) Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

6.34 Barking deer. Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests make up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.[45]

Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	61	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

(b) What type of test can we use to answer this research question?

(c) Check if the assumptions and conditions required for this test are satisfied.

(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

Photo by Shrikant Rao
(http://flic.kr/p/4Xjdkk)
CC BY 2.0 license

---

[45]Liwei Teng et al. "Forage and bed sites characteristics of Indian muntjac (Muntiacus muntjak) in Hainan Island, China". In: *Ecological Research* 19.6 (2004), pp. 675–681.

6.4   Chi-square tests for two-way tables

We encounter two-way tables in this section, and we learn about two new and closely related chi-square tests. We will answer questions such as the following:

- Does the phrasing of the question affect how likely sellers are to disclose problems with a product?

- Is gender associated with whether Facebook users know how to adjust their privacy settings?

- Is political affiliation associated with support for the use of full body scans at airports?

Learning objectives

1. Calculate the expected counts and degrees of freedom for a chi-square test involving a two-way table.

2. State and verify whether or not the conditions for a chi-square test for a two-way table are met.

3. Explain the difference between the chi-square test of homogeneity and chi-square test of independence.

4. Carry out a complete hypothesis test for homogeneity and for independence.

6.4.1 Introduction

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Figure 6.10 shows an example of 10,000 queries split into three algorithm groups.[46] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

Figure 6.10: Experiment breakdown of test subjects into three search groups.

EXAMPLE 6.42

What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

H_0: The algorithms each perform equally well.

H_A: The algorithms do not perform equally well.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reflect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) there was no new, related search, and the user clicked one of the links provided, or (2) there was a new, related search performed by the user. Under scenario (1), we might think that the user was satisfied with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Figure 6.11 provides the results from the experiment. These data are very similar to the count data in Section 6.3. However, now the different combinations of two variables are binned in a *two-way* table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

	Search algorithm			
	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Figure 6.11: Results of the Google search algorithm experiment.

WHAT IS SO DIFFERENT ABOUT ONE-WAY TABLES AND TWO-WAY TABLES?

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for *combinations* of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the three search algorithms perform differently.

---

[46]Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

6.4.2   Expected counts in two-way tables

EXAMPLE 6.43

From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as $7078/10000 = 0.7078$. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satisfied with the initial search:

$$0.7078 \times 5000 = 3539 \text{ users}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

GUIDED PRACTICE 6.44

Using the same rationale described in Example 6.43, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[47]

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 6.43 and Guided Practice 6.44. These expected counts were used to construct Figure 6.12, which is the same as Figure 6.11, except now the expected counts have been added in parentheses.

Search algorithm	current		test 1		test 2		Total
No new search	3511	**(3539)**	1749	**(1769.5)**	1818	**(1769.5)**	7078
New search	1489	**(1461)**	751	**(730.5)**	682	**(730.5)**	2922
Total	5000		2500		2500		10000

Figure 6.12: The observed counts and the **(expected counts)**.

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the first row:

$$0.7078 \times (\text{column 1 total}) = 3539$$
$$0.7078 \times (\text{column 2 total}) = 1769.5$$
$$0.7078 \times (\text{column 3 total}) = 1769.5$$

Looking back to how the fraction 0.7078 was computed – as the fraction of users who did not perform a new search $(7078/10000)$ – these three expected counts could have been computed as

$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) = 3539$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) = 1769.5$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) = 1769.5$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

---

[47]We would expect $0.7078 * 2500 = 1769.5$. It is okay that this is a fraction.

COMPUTING EXPECTED COUNTS IN A TWO-WAY TABLE

To identify the expected count for the ith row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

6.4.3 The chi-square test of homogeneity for two-way tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

General formula $\qquad \dfrac{(\text{observed count } - \text{ expected count})^2}{\text{expected count}}$

Row 1, Col 1 $\qquad \dfrac{(3511 - 3539)^2}{3539} = 0.222$

Row 1, Col 2 $\qquad \dfrac{(1749 - 1769.5)^2}{1769.5} = 0.237$

$\qquad\qquad\vdots \qquad\qquad\qquad\qquad \vdots$

Row 2, Col 3 $\qquad \dfrac{(682 - 730.5)^2}{730.5} = 3.220$

Adding the computed value for each cell gives the chi-square test statistic χ^2:

$$\chi^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom is computed a little differently for a two-way table.[48] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows - 1}) \times (\text{number of columns - 1})$$

In our example, the degrees of freedom is

$$df = (2 - 1) \times (3 - 1) = 2$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic $\chi^2 = 6.12$ closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.13.

COMPUTING DEGREES OF FREEDOM FOR A TWO-WAY TABLE

When using the chi-square test to a two-way table, we use

$$df = (R - 1) \times (C - 1)$$

where R is the number of rows in the table and C is the number of columns.

USE TWO-PROPORTION METHODS FOR 2-BY-2 CONTINGENCY TABLES

When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 6.2.

---

[48]Recall: in the one-way table, the degrees of freedom was the number of groups minus 1.
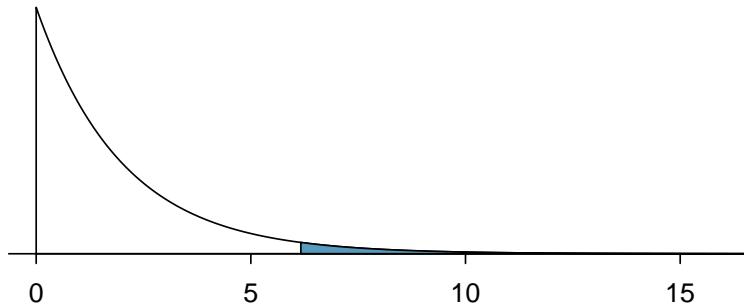
Figure 6.13: Computing the p-value for the Google hypothesis test.

CONDITIONS FOR THE CHI-SQUARE TEST OF HOMOGENEITY

There are two conditions that must be checked before performing a chi-square test of homogeneity. If these conditions are not met, this test should not be used.

Multiple random samples or randomly assigned treatments. Data collected by multiple independent random samples or multiple randomly assigned treatments. Data can then be organized into a two-way table.

All expected counts at least 5. All of the cells in the two-way table must have at least 5 expected cases under the assumption that the null hypothesis is true.

EXAMPLE 6.45

Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

———

Here, found that the degrees of freedom for this 3×2 table is 2. The p-value corresponds to the area under the chi-square curve with 2 degrees of freedom to the *right* of $\chi^2 = 6.120$. Using a calculator, we find that the p-value = 0.047. Using an $\alpha = 0.05$ significance level, we reject H_0. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Notice that the conclusion of the test is that there is some difference in performance among the algorithms. This chi-square test does not tell us *which* algorithm performed better than the others. To answer this question, we could compare the relevant proportions or construct bar graphs. The proportion that resulted in the new search can be calculated as

$$\text{current:}\frac{1489}{5000} = 0.298 \quad \text{test 1:}\frac{751}{2500} = 0.300 \quad \text{test 2:}\frac{682}{2500} = 0.136.$$

This suggests that the current algorithm and test 1 algorithm performed better than the test 2 algorithm; however, to formally test this specific claim we would need to use a test that includes a multiple comparisons correction, which is beyond the scope of this textbook.

A careful reader may have noticed that when there are exactly 2 random samples or treatments and the counts can be arranged in a 2×2 table, both a chi-square test for homogeneity *and* a 2-proportion Z-test could apply. In this case, the chi-square test for homogeneity and the two-sided 2-proportion Z-test are equivalent, meaning that they produce the same p-value.[49]

---

[49]Sometimes the success-failure condition for the Z-test is weakened to require the number of successes and failures to be at least 5, making it consistent with the chi-square condition that expected counts must at least 5.

χ^2 TEST OF HOMOGENEITY

When there are multiple samples or treatments and we are comparing the distribution of a categorical variable across several groups, e.g. comparing the distribution of rural/urban/suburban dwellers among 4 states,

Identify: Identify the hypotheses and the significance level, α.

H_0: The distribution of [...] is the same for each population/treatment.

H_A: The distribution of [...] is not the same for each population/treatment.

Choose: Choose the correct test procedure and identify it by name.

Here we choose the $\boldsymbol{\chi^2}$ **test of homogeneity**.

Check: Check that the test statistic follows a chi-square distribution.

1. Data come from multiple random samples or from multiple randomly assigned treatments.

2. All expected counts are ≥ 5 (calculate and record expected counts).

Calculate: Calculate the χ^2-statistic, df, and p-value.

test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

$df = (\# \text{ of rows} - 1) \times (\# \text{ of columns} - 1)$

p-value = (area to the *right* of χ^2-statistic with the appropriate df)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].

EXAMPLE 6.46

In an experiment,[50] each individual was asked to be a seller of an iPod (a product commonly used to store music on before smart phones). The participant received $10 + 5\%$ of the sale price for participating. The iPod they were selling had frozen twice in the past inexplicitly but otherwise worked fine. Unbeknownst to the participants who were the sellers in the study, the buyers were collaborating with the researchers to evaluate the influence of different questions on the likelihood of getting the sellers to disclose the past issues with the iPod. The scripted buyers started with "Okay, I guess I'm supposed to go first. So you've had the iPod for 2 years ..." and ended with one of three questions:

- General: What can you tell me about it?

- Positive Assumption: It doesn't have any problems, does it?

- Negative Assumption: What problems does it have?

The outcome variable is whether the participant discloses or hides the problem with the iPod.

		Question Type		
		General	Positive Assump.	Negative Assump.
Response	Disclose	2	23	36
	Hide	71	50	37
	Total	73	73	73

Does the phrasing of the question affect how likely individuals are to disclose the problems with the iPod? Carry out an appropriate test at the 0.05 significance level.

Ⓔ

Identify: We will test the following hypotheses at the $\alpha = 0.05$ significance level.
H_0: The likelihood of disclosing the problem is the same for each question type.
H_A: The likelihood of disclosing the problem is not the same for each question type.

Choose: We want to know if the distribution of disclose/hide is the same for each of the three question types, so we want to carry out a chi-square test for homogeneity.

Check: This is an experiment in which there were three randomly allocated treatments. Here a treatment corresponds to a question type. All values in the table of expected counts are ≥ 5. Table of expected counts:

		Question Type		
		General	Positive Assump.	Negative Assump.
Response	Disclose	20.3	20.3	20.3
	Hide	52.7	52.7	52.7

Calculate: Using technology, we get $\chi^2 = 40.1$
$df = (\# \text{ of rows} - 1) \times (\# \text{ of columns} - 1) = 2 \times 1 = 2$
The p-value is the area under the chi-square curve with 2 degrees of freedom to the right of $\chi^2 = 40.1$. Thus, the p-value is almost 0.

Conclude: Because the p-value $\approx 0 < \alpha$, we reject H_0. We have strong evidence that the likelihood of disclosing the problem is not the same for each question type.

GUIDED PRACTICE 6.47

Ⓖ

If an error was made in the test in the previous example, would it have been a Type I error or a Type II error?[51]

---

[50]opim.wharton.upenn.edu/DPlab/papers/workingPapers/
[51]In this test, the p-value was less than α, so we rejected H_0. If H_0 is in fact true, and we reject it, that would be committing a Type I error. We could not have made a Type II error, because a Type II error involves not rejecting H_0.

6.4.4 The chi-square test of independence for two-way tables

Often, instead of having separate random samples or treatments, we have just one sample and we want to look at the association between two variables. When these two variables are categorical, we can arrange the responses in a two-way table.

In Chapter 3 we looked at independence in the context of probability. Here we look at independence in the context of inference. We want to know if any observed association is due to random chance or if there is evidence of a real association in the population that the sample was taken from. To answer this, we use a chi-square test for independence. The chi-square test of independence applies when there is only one random sample and there are two categorical variables. The null claim is always that the two variables are independent, while the alternate claim is that the variables are dependent.

EXAMPLE 6.48

Figure 6.14 summarizes the results of a Pew Research poll.[52] A random sample of adults in the U.S. was taken, and each was asked whether they approved or disapproved of the job being done by President Obama, Democrats in Congress, and Republicans in Congress. The results are shown in Figure 6.14. We would like to determine if the three groups and the approval ratings are associated. What are appropriate hypotheses for such a test?

H_0: The group and their ratings are independent. (There is no difference in approval ratings between the three groups.)

H_A: The group and their ratings are dependent. (There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.)

		Congress		
	Obama	Democrats	Republicans	Total
Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

Figure 6.14: Pew Research poll results of a March 2012 poll.

CONDITIONS FOR THE CHI-SQUARE TEST OF INDEPENDENCE

There are two conditions that must be checked before performing a chi-square test of independence. If these conditions are not met, this test should not be used.

One random sample with two variables/questions. The data must be arrived at by taking a random sample. After the data is collected, it is separated and categorized according to two variables and can be organized into a two-way table.

All expected counts at least 5. All of the cells in the two-way table must have at least 5 expected cases assuming the null hypothesis is true.

---

[52]www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama.

EXAMPLE 6.49

First, we observe that the data came from a random sample of adults in the U.S. Next, let's compute the expected values that correspond to Figure 6.14, if the null hypothesis is true, that is, if group and rating are independent.

The expected count for row one, column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119 \times 1458}{4223} = 731.6$. Similarly for the first column and the second row: $\frac{2104 \times 1458}{4223} = 726.4$. Repeating this process, we get the expected counts:

	Obama	Congr. Dem.	Congr. Rep.
Approve	731.6	693.5	694.0
Disapprove	726.4	688.5	689.0

The table above gives us the number we would expect for each of the six combinations if group and rating were really independent. Because all of the expected counts are at least 5 and there is one random sample, we can carry out the chi-square test for independence.

The chi-square test of independence and the chi-square test of homogeneity both involve counts in a two-way table. The chi-square statistic and the degrees of freedom are calculated in the same way.

EXAMPLE 6.50

Calculate the chi-square statistic.

We calculate $\frac{(\text{obs}-\text{exp})^2}{\text{exp}}$ for each of the six cells in the table. Adding the results of each cell gives the chi-square test statistic.

$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$
$$= \frac{(842 - 731.6)^2}{731.6} + \cdots$$
$$= 16.7 + \cdots = 106.4$$

EXAMPLE 6.51

Find the p-value for the test and state the appropriate conclusion.

We must first find the degrees of freedom for this chi-square test. Because there are 2 rows and 3 columns, the degrees of freedom is $df = (2 - 1) \times (3 - 1) = 2$. We find the area to the right of $\chi^2 = 106.4$ under the chi-square curve with $df = 2$. The p-value is extremely small, much less than 0.01, so we reject H_0. We have evidence that the three groups and their approval ratings are dependent.

χ^2 **TEST OF INDEPENDENCE**

When there is one sample and we are looking for association or dependence between two categorical variables, e.g. testing for an association between gender and political party,

Identify: Identify the hypotheses and the significance level, α.

H_0: [variable 1] and [variable 2] are independent.

H_A: [variable 1] and [variable 2] are dependent.

Choose: Choose the correct test procedure and identify it by name.

Here we choose the χ^2 **test of independence**.

Check: Check that the test statistic follows a chi-square distribution.

1. Data come from a single random sample.

2. All expected counts are ≥ 5 (calculate and record expected counts).

Calculate: Calculate the χ^2-statistic, df, and p-value.

test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

$df = (\# \text{ of rows} - 1) \times (\# \text{ of columns} - 1)$

p-value = (area to the *right* of χ^2-statistic with the appropriate df)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that $[H_A$ in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that $[H_A$ in context].

EXAMPLE 6.52

A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, "Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?" The responses are cross-tabulated based on gender.

| | | *Gender* |
		Male	Female	Total
Response	Yes	288	378	666
	No	61	62	123
	Not sure	10	7	17
	Total	359	447	806

Carry out an appropriate test at the 0.10 significance level to see if there is an association between gender and knowing how to adjust Facebook privacy settings to control what people can and cannot see.

—————

Identify: We will test the following hypotheses at the $\alpha = 0.10$ significance level.
   H_0: Gender and knowing how to adjust Facebook privacy settings are independent.
   H_A: Gender and knowing how to adjust Facebook privacy settings are dependent.

Choose: Two variables were recorded on the respondents: gender and response to the question regarding privacy settings. We want to know if these variables are associated / dependent, so we will carry out a chi-square test of independence.

Check: According to the problem, there was one random sample taken. All values in the table of expected counts are ≥ 5. Table of expected counts:

| | | *Gender* |
		Male	Female
Response	Yes	296.64	369.36
	No	54.785	68.215
	Not sure	7.572	9.428

Calculate: Using technology, we get $\chi^2 = 3.13$. The degrees of freedom for this test is given by:
   $df = (\text{\# of rows} - 1) \times (\text{\# of columns} - 1) = 2 \times 1 = 2$
   The p-value is the area under the chi-square curve with 2 degrees of freedom to the right of $\chi^2 = 3.13$. Thus, the p-value = 0.209.

Conclude: Because the p-value = $0.209 > \alpha$, we do not reject H_0. We do not have sufficient evidence that gender and knowing how to adjust Facebook privacy settings are dependent.

GUIDED PRACTICE 6.53

In context, interpret the p-value of the test in the previous example.[53]

———————————

[53]The p-value in this test corresponds to the area to the right of $\chi^2 = 3.13$ under the chi-square curve with 2 degrees of freedom. It is the probability of getting a χ^2-statistic larger than 3.13 if H_0 were true and assuming a chi-square model with 2 degrees of freedom holds. Equivalently, it is the probability of our observed counts being this different from the expected counts, relative to the expected counts, if gender and response really are independent (and the model holds).

6.4.5 Calculator: chi-square test for two-way tables

TI-83/84: ENTERING DATA INTO A TWO-WAY TABLE

1. Hit `2ND x`$^{-1}$ (i.e. `MATRIX`).
2. Right arrow to `EDIT`.
3. Hit `1` or `ENTER` to select matrix `A`.
4. Enter the dimensions by typing #rows, `ENTER`, #columns, `ENTER`.
5. Enter the data from the two-way table.

TI-83/84: CHI-SQUARE TEST OF HOMOGENEITY AND INDEPENDENCE

Use `STAT`, `TESTS`, χ^2`-Test`.

1. First enter two-way table data as described in the previous box.
2. Choose `STAT`.
3. Right arrow to `TESTS`.
4. Down arrow and choose `C:`χ^2`-Test`.
5. Down arrow, choose `Calculate`, and hit `ENTER`, which returns

χ^2	chi-square test statistic
p	p-value
df	degrees of freedom

TI-83/84: CHI-SQUARE TEST OF HOMOGENEITY AND INDEPENDENCE

TI-83/84: Finding the expected counts

1. First enter two-way table data as described previously.
2. Carry out the chi-square test of homogeneity or independence as described in previous box.
3. Hit `2ND x`$^{-1}$ (i.e. `MATRIX`).
4. Right arrow to `EDIT`.
5. Hit `2` to see matrix `B`.

   This matrix contains the expected counts.

CASIO FX-9750GII: CHI-SQUARE TEST OF HOMOGENEITY AND INDEPENDENCE

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).
2. Choose the `TEST` option (`F3` button).
3. Choose the `CHI` option (`F3` button).
4. Choose the `2WAY` option (`F2` button).
5. Enter the data into a matrix:

   - Hit ▷`MAT` (`F2` button).
   - Navigate to a matrix you would like to use (e.g. `Mat C`) and hit `EXE`.
   - Specify the matrix dimensions: `m` is for rows, `n` is for columns.
   - Enter the data.
   - Return to the test page by hitting `EXIT` twice.

6. Enter the `Observed` matrix that was used by hitting `MAT` (`F1` button) and the matrix letter (e.g. `C`).
7. Enter the `Expected` matrix where the expected values will be stored (e.g. `D`).
8. Hit the `EXE` button, which returns

   χ^2 chi-square test statistic
   `p` p-value
   `df` degrees of freedom

9. To see the expected values of the matrix, go to ▷`MAT` (`F6` button) and select the corresponding matrix.

GUIDED PRACTICE 6.54

Use Figure 6.14, reproduced below, and a calculator to find the expected values and the χ^2-statistic, df, and p-value for the chi-square test for independence.[54]

Ⓖ

		Congress		
	Obama	Democrats	Republicans	Total
Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

---

[54]First create a 2×3 matrix with the data. The final summaries should be $\chi^2 = 106.4$, p-value is $p = 8.06 \times 10^{-24} \approx 0$, and `df` $= 2$. Below is the matrix of expected values:

	Obama	Congr. Dem.	Congr. Rep.
Approve	731.59	693.45	693.96
Disapprove	726.41	688.55	689.04

Section summary

- When there are two categorical variables, rather than one, the data must be arranged in a **two-way table** and a χ^2 test of homogeneity or a χ^2 test of independence is appropriate.

- These tests use the same χ^2-statistic as the chi-square goodness of fit test, but instead of number of categories $- 1$, the **degrees of freedom** is (# of rows $- 1$) $\times$ (# of columns $- 1$). All expected counts must be at least 5.

- When working with a two-way table, the **expected count** for each row,column combination is calculated as: expected count $= \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$.

- The χ^2 test of homogeneity and the χ^2 test of independence are almost identical. The differences lie in the data collection method and in the hypotheses.

- When there are **multiple samples or treatments** and we are comparing the distribution of a categorical variable across several groups, e.g. comparing the distribution of rural/urban/suburban dwellers among 4 states, the hypotheses can often be written as follows:

  H_0: The distribution of [...] is the same for each population/treatment.

  H_A: The distribution of [...] is not the same for each population/treatment.

  We test these hypotheses at the α significance level using a **χ^2 test of homogeneity**.

- When there is **one sample** and we are looking for association or dependence between two categorical variables, e.g. testing for an association between gender and political party, the hypotheses can be written as:

  H_0: [variable 1] and [variable 2] are independent.

  H_A: [variable 1] and [variable 2] are dependent.

  We test these hypotheses at the α significance level using a **χ^2 test of independence**.

- Both of the χ^2 tests for two-way tables require that all expected counts are ≥ 5.

- The chi-square statistic is:

  test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

  $df = (\text{\# of rows} - 1)(\text{\# of cols} - 1)$

- The p-value is the area to the *right* of χ^2-statistic under the chi-square curve with the appropriate df.

Exercises

6.35   Quitters. Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

(a) Create a two-way table presenting the results of this study.

(b) Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

   i. How many subjects in the "patch + support" group would you expect to quit?
   ii. How many subjects in the "patch only" group would you expect to not quit?

6.36   Full body scan, Part II. The table below summarizes a data set we first encountered in Exercise 6.26 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's party affiliation and his support of full-body scans. It may be useful to first add on an extra column for row totals before proceeding with the computations.

		Party Affiliation		
		Republican	Democrat	Independent
	Should	264	299	351
Answer	Should not	38	55	77
	Don't know/No answer	16	15	22
	Total	318	369	450

(a) How many Republicans would you expect to not support the use of full-body scans?

(b) How many Democrats would you expect to support the use of full- body scans?

(c) How many Independents would you expect to not know or not answer?

6.37   Offshore drilling, Part III. 🎥   The table below summarizes a data set we first encountered in Exercise 6.23 that examines the responses of a random sample of college graduates and non-graduates on the topic of oil drilling. Complete a chi-square test for these data to check whether there is a statistically significant difference in responses from college graduates and non-graduates.

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

6.38   Parasitic worm. Lymphatic filariasis is a disease caused by a parasitic worm. Complications of the disease can lead to extreme swelling and other complications. Here we consider results from a randomized experiment that compared three different drug treatment options to clear people of the this parasite, which people are working to eliminate entirely. The results for the second year of the study are given below:[55]

	Clear at Year 2	Not Clear at Year 2
Three drugs	52	2
Two drugs	31	24
Two drugs annually	42	14

(a) Set up hypotheses for evaluating whether there is any difference in the performance of the treatments, and also check conditions.

(b) Statistical software was used to run a chi-square test, which output:

$$X^2 = 23.7 \qquad\qquad df = 2 \qquad\qquad \text{p-value} = 7.2\text{e-}6$$

Use these results to evaluate the hypotheses from part (a), and provide a conclusion in the context of the problem.

---

[55]Christopher King et al. "A Trial of a Triple-Drug Treatment for Lymphatic Filariasis". In: *New England Journal of Medicine* 379 (2018), pp. 1801–1810.

Chapter highlights

Calculating a confidence interval or a test statistic and p-value are generally done with statistical software. It is important, then, to focus not on the calculations, but rather on

1. choosing the correct procedure
2. understanding when the procedures do or do not apply, and
3. interpreting the results.

Choosing the correct procedure requires understanding the *type of data* and the *method of data collection*. All of the inference procedures in Chapter 6 are for categorical variables. Here we list the five tests encountered in this chapter and when to use them.

- **1-proportion Z-test**
    - *1 random sample*, a yes/no variable
    - Compare the sample proportion to a fixed / hypothesized proportion.

- **2-proportion Z-test**
    - *2 independent random samples or randomly allocated treatments*
    - Compare two populations or treatments to each other with respect to one yes/no variable; e.g. comparing the proportion over age 65 in two distinct populations.

- **χ^2 goodness of fit test**
    - *1 random sample*, a categorical variable (generally at least three categories)
    - Compare the distribution of a categorical variable to a fixed or known population distribution; e.g. looking at distribution of color among M&M's.

- **χ^2 test of homogeneity**:
    - *2 or more independent random samples or randomly allocated treatments*
    - Compare the distribution of a categorical variable across several populations or treatments; e.g. party affiliation over various years, or patient improvement compared over 3 treatments.

- **χ^2 test of independence**
    - *1 random sample*, 2 categorical variables
    - Determine if, in a single population, there is an association between two categorical variables; e.g. grade level and favorite class.

Even when the data and data collection method correspond to a particular test, we must *verify that conditions are met* to see if the assumptions of the test are reasonable. All of the inferential procedures of this chapter require some type of random sample or process. In addition, the 1-proportion Z-test/interval and the 2-proportion Z-test/interval require that the success-failure condition is met and the three χ^2 tests require that all expected counts are at least 5.

Finally, understanding and communicating the logic of a test and being able to accurately *interpret* a confidence interval or p-value are essential. For a refresher on this, review Chapter 5: Foundations for inference.

Chapter exercises

6.39   Active learning. A teacher wanting to increase the active learning component of her course is concerned about student reactions to changes she is planning to make. She conducts a survey in her class, asking students whether they believe more active learning in the classroom (hands on exercises) instead of traditional lecture will helps improve their learning. She does this at the beginning and end of the semester and wants to evaluate whether students' opinions have changed over the semester. Can she used the methods we learned in this chapter for this analysis? Explain your reasoning.

6.40   Website experiment. The OpenIntro website occasionally experiments with design and link placement. We conducted one experiment testing three different placements of a download link for this textbook on the book's main page to see which location, if any, led to the most downloads. The number of site visitors included in the experiment was 701 and is captured in one of the response combinations in the following table:

	Download	No Download
Position 1	13.8%	18.3%
Position 2	14.6%	18.5%
Position 3	12.1%	22.7%

(a) Calculate the actual number of site visitors in each of the six response categories.

(b) Each individual in the experiment had an equal chance of being in any of the three experiment groups. However, we see that there are slightly different totals for the groups. Is there any evidence that the groups were actually imbalanced? Make sure to clearly state hypotheses, check conditions, calculate the appropriate test statistic and the p-value, and make your conclusion in context of the data.

(c) Complete an appropriate hypothesis test to check whether there is evidence that there is a higher rate of site visitors clicking on the textbook link in any of the three groups.

6.41   Shipping holiday gifts. A local news survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age			
		18-34	35-54	55+	Total
	USPS	72 (81)	97 (102)	76 (62)	245
	UPS	52 (53)	76 (68)	34 (41)	162
Shipping Method	FedEx	31 (21)	24 (27)	9 (16)	64
	Something else	7 (5)	6 (7)	3 (4)	16
	Not sure	3 (5)	6 (5)	4 (3)	13
	Total	165	209	126	500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

(b) Are the conditions for inference using a chi-square test satisfied?

6.42   The Civil War. A national survey conducted among a simple random sample of 1,507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life.[56]

(a) Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) Interpret the p-value in this context.

(c) Calculate a 90% confidence interval for the proportion of Americans who think the Civil War is still relevant. Interpret the interval in this context, and comment on whether or not the confidence interval agrees with the conclusion of the hypothesis test.

---

[56]Pew Research Center Publications, *Civil War at 150: Still Relevant, Still Divisive*, data collected between March 30 - April 3, 2011.

6.43 College smokers. We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke.

(a) Calculate a 95% confidence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: Check conditions.)

(b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?

6.44 Acetaminophen and liver damage. It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher wants to conduct a study to estimate the proportion of acetaminophen users who have liver damage. For participating in this study, he will pay each subject $20 and provide a free medical consultation if the patient has liver damage.

(a) If he wants to limit the margin of error of his 98% confidence interval to 2%, what is the minimum amount of money he needs to set aside to pay his subjects?

(b) The amount you calculated in part (a) is substantially over his budget so he decides to use fewer subjects. How will this affect the width of his confidence interval?

6.45 Life after college. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions for constructing a confidence interval based on these data are met.

(c) Calculate a 95% confidence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

(d) What does "95% confidence" mean?

(e) Now calculate a 99% confidence interval for the same parameter and interpret it in the context of the data.

(f) Compare the widths of the 95% and 99% confidence intervals. Which one is wider? Explain.

6.46 Diabetes and unemployment. A Gallup poll surveyed Americans about their employment status and whether or not they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.[57]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for difference in proportions of diabetes between employed and unemployed Americans.

(c) The sample difference is about 1%. If we completed the hypothesis test, we would find that the p-value is very small (about 0), meaning the difference is statistically significant. Use this result to explain the difference between statistically significant and practically significant findings.

6.47 Rock-paper-scissors. Rock-paper-scissors is a hand game played by two or more people where players choose to sign either rock, paper, or scissors with their hands. For your statistics class project, you want to evaluate whether players choose between these three options randomly, or if certain options are favored above others. You ask two friends to play rock-paper-scissors and count the times each option is played. The following table summarizes the data:

Rock	Paper	Scissors
43	21	35

Use these data to evaluate whether players choose between these three options randomly, or if certain options are favored above others. Make sure to clearly outline each step of your analysis, and interpret your results in context of the data and the research question.

---

[57]Gallup Wellbeing, *Employed Americans in Better Health Than the Unemployed*, data collected Jan. 2, 2011 - May 21, 2012.

6.48   2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[58]

(a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

(d) The margin of error at a 90% confidence level would be higher than 3%.

6.49   Browsing on the mobile device. A survey of 2,254 American adults indicates that 17% of cell phone owners browse the internet exclusively on their phone rather than a computer or other device.[59]

(a) According to an online article, a report from a mobile research company indicates that 38 percent of Chinese mobile web users only access the internet through their cell phones.[60] Conduct a hypothesis test to determine if these data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%.

(b) Interpret the p-value in this context.

(c) Calculate a 95% confidence interval for the proportion of Americans who access the internet on their cell phones, and interpret the interval in this context.

6.50   Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician- diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[61]

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical	Yes	670	373	905	564	95	2,607
depression	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identified in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$.

(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.[62] Do you agree with this statement? Explain your reasoning.

[58]Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.

[59]Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.

[60]S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: *M.I.C Gadget* (2012).

[61]M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: *Archives of internal medicine* 171.17 (2011), p. 1571.

[62]A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: *New York Times* (2011).

Chapter 7

Inference for numerical data

Chapter 5 introduced a framework for statistical inference based on confidence intervals and hypothesis tests. Chapter 6 summarized inference procedures for categorical data (counts and proportions), using the normal distribution and the chi-square distribution. In this chapter, we focus on inference procedures for numerical data and we encounter a new distribution. In each case, the inference ideas remain the same:

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Apply the ideas from Chapter 5 using the distribution from step 2.

Each section in Chapter 7 explores a new situation: a single mean (7.1), a mean of differences (7.2); and a difference of means (7.3).

For videos, slides, and other resources, please visit
www.openintro.org/ahss

7.1 Inference for a mean with the t-distribution

In this section, we turn our attention to numerical variables and answer questions such as the following:

- How well can we estimate the mean income of people in a certain city, county, or state?

- What is the average mercury content in various types of fish?

- Are people's run times getting faster or slower, on average?

- How does the sample size affect the expected error in our estimates?

- When is it reasonable to model the sample mean $\bar{x}$ using a normal distribution, and when will we need to use a new distribution, known as the t-distribution?

Learning objectives

1. Understand the relationship between a t-distribution and a normal distribution, and explain why we use a t-distribution for inference on a mean.

2. State and verify whether or not the conditions for inference for a mean based on the t-distribution are met. Understand when it is necessary to look at the distribution of the sample data.

3. Know the degrees of freedom associated with a one sample t-procedure.

4. Carry out a complete hypothesis test for a single mean.

5. Carry out a complete confidence interval procedure for a single mean.

6. Find the minimum sample size needed to estimate a mean with C% confidence and a margin of error no greater than a certain value.

7.1.1 Using a normal distribution for inference when σ is known

In Section 4.2 we saw that the distribution of a sample mean is normal if the population is normal or if the sample size is at least 30. In these problems, we used the population mean and population standard deviation to find a Z-score. However, in the case of inference, these values will be unknown. In rare circumstances we may know the standard deviation of a population, even though we do not know its mean. For example, in some industrial processes, the mean may be known to shift over time, while the standard deviation of the process remains the same. In these cases, we can use the normal model as the basis for our inference procedures. We use $\bar{x}$ as our point estimate for μ and the SD formula for a sample mean calculated in Section 4.2: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. That leads to a confidence interval and a test statistic as follows:

$$\text{CI: } \bar{x} \ \pm \ z^{\star}\frac{\sigma}{\sqrt{n}} \qquad\qquad Z = \frac{\bar{x} - \text{null value}}{\frac{\sigma}{\sqrt{n}}}$$

What happens if we do not know the population standard deviation σ, as is usually the case? The best we can do is use the sample standard deviation, denoted by s, to estimate the population standard deviation.

$$SE = \frac{s}{\sqrt{n}}$$

However, when we do this we run into a problem: when carrying out our inference procedures, we will be trying to estimate *two* quantities: both the mean and the standard deviation. Looking at the SD and SE formulas, we can make some important observations that will give us a hint as to what will happen when we use s instead of σ.

- For a given population, σ is a fixed number and does not vary.
- s, the standard deviation of a sample, will vary from one sample to the next and will not be exactly equal to σ.
- The larger the sample size n, the better the estimate s will tend to be for σ.

For this reason, the normal model still works well when the sample size is large. For smaller sample sizes, we run into a problem: our use of s, which is used when computing the standard error, tends to add more variability to our test statistic. It is this extra variability that leads us to a new distribution: the t-distribution.

7.1.2   Introducing the t-distribution

When we use the sample standard deviation s in place of the population standard deviation σ to standardize the sample mean, we get an entirely new distribution - one that is similar to the normal distribution, but has greater spread. This distribution is known as the t-distribution. A t-distribution, shown as a solid line in Figure 7.1, has a bell shape. However, its tails are thicker than the normal model's. We can see that a greater proportion of the area under the t-distribution is beyond 2 standard units from 0 than under the normal distribution. These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard deviation.
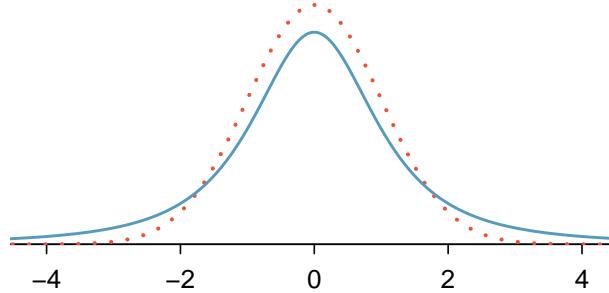
Figure 7.1: Comparison of a t-distribution (solid line) and a normal distribution (dotted line).

The t-distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describes the precise form of the bell-shaped t-distribution. Several t-distributions are shown in Figure 7.2. When there are more degrees of freedom, the t-distribution looks more like the standard normal distribution.

DEGREES OF FREEDOM

The degrees of freedom describes the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution resembles the standard normal distribution.

When the degrees of freedom is large, about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution. In Section 7.1.4, we will see how degrees of freedom relates to sample size.
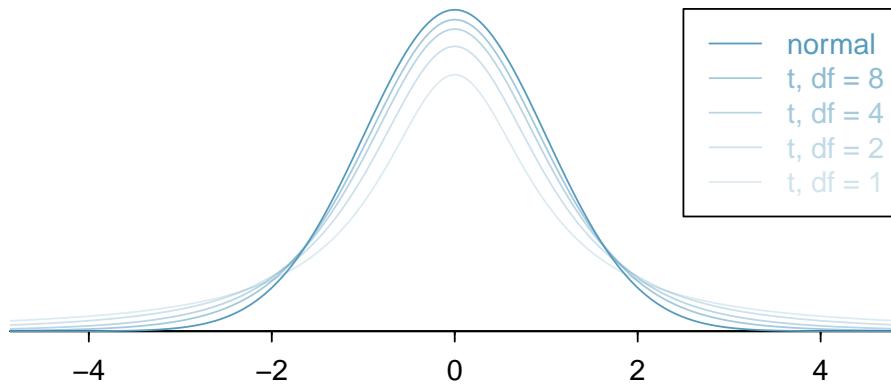
Figure 7.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal distribution.

We will find it useful to become familiar with the t-distribution, because it plays a very similar role to the normal distribution during inference. We use a **t-table**, partially shown in Figure 7.3, in place of the normal probability table when the population standard deviation is unknown, especially when the sample size is small. A larger table is presented in Appendix C.3.

	one tail	0.100	0.050	0.025	0.010	0.005
df	1	3.078	6.314	12.71	31.82	63.66
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	⋮	⋮	⋮	⋮	⋮	⋮
	17	1.333	1.740	2.110	2.567	2.898
	18	**1.330**	**1.734**	**2.101**	**2.552**	**2.878**
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	⋮	⋮	⋮	⋮	⋮	⋮
	1000	1.282	1.646	1.962	2.330	2.581
	∞	1.282	1.645	1.960	2.326	2.576
Confidence level C		80%	90%	95%	98%	99%

Figure 7.3: An abbreviated look at the t-table. Each row represents a different t-distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been highlighted.

Each row in the t-table represents a t-distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Figure 7.3. If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

EXAMPLE 7.1

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw a picture as shown in Figure 7.4 and shade the area below -2.10. To find this area, we identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. That is, 2.5% of the distribution falls below -2.10.
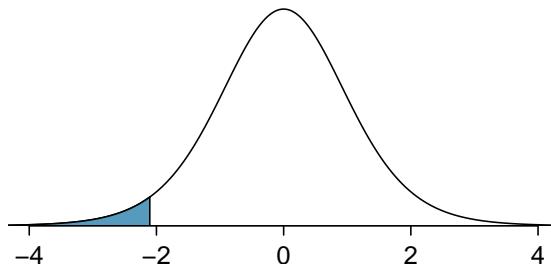
Figure 7.4: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.
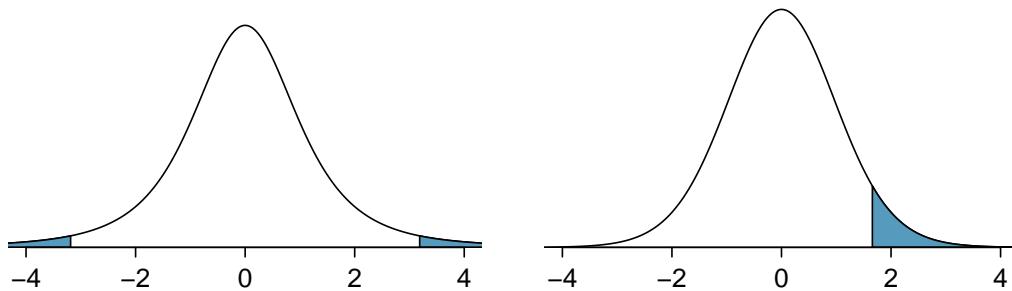
Figure 7.5: Left: The t-distribution with 3 degrees of freedom, with the area farther than 3.182 units from 0 shaded. Right: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded.

EXAMPLE 7.2

For the t-distribution with 18 degrees of freedom, what percent of the curve is contained between -1.330 and +1.330?

Using row $df = 18$, we find 1.330 in the table. The area in each tail is 0.100 for a total of 0.200, which leaves 0.800 in the middle between -1.33 and +1.33. This corresponds to the 80%, which can be found at the very bottom of that column.

EXAMPLE 7.3

For the t-distribution with 3 degrees of freedom, as shown in the left panel of Figure 7.5, what should the value of $t^\star$ be so that 95% of the area of the curve falls between $-t^\star$ and $+t^\star$?

We can look at the column in the t-table that says 95% along the bottom row and trace it up to row $df = 3$ to find that $t^\star = 3.182$.

EXAMPLE 7.4

A t-distribution with 20 degrees of freedom is shown in the right panel of Figure 7.5. Estimate the proportion of the distribution falling above 1.65.

We identify the row in the t-table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

When the desired degrees of freedom is not listed on the table, choose a conservative value: round the degrees of freedom down, i.e. move *up* to the previous row listed. Another option is to use a calculator or statistical software to get a precise answer.

7.1.3 Calculator: finding area under the t-distribution

It is possible to find areas under a t-distribution on a calculator.

TI-84: FINDING AREA UNDER THE T-CURVE

Use 2ND VARS, tcdf to find an area/proportion/probability between two t-scores or to the left or right of a t-score.

1. Choose 2ND VARS (i.e. DISTR).
2. Choose 6:tcdf.
3. Enter the lower (left) t-score and the upper (right) t-score.
   - If finding just a lower tail area, set lower to -100.
   - For an upper tail area, set upper to 100.
4. Enter the degrees of freedom after df:.
5. Down arrow, choose Paste, and hit ENTER.

TI-83: Do steps 1-2, then enter the lower bound, upper bound, degrees of freedom, e.g. tcdf(2, 100, 5), and hit ENTER.

CASIO FX-9750GII: FINDING AREA UNDER THE T-DISTRIBUTION

1. Navigate to STAT (MENU, then hit 2).
2. Select DIST (F5), then t (F2), and then tcd (F2).
3. If needed, set Data to Variable (Var option, which is F2).
4. Enter the Lower t-score and the Upper t-score. Set the degrees of freedom (df).
   - If finding just a lower tail area, set Lower to -100.
   - For an upper tail area, set Upper to 100.
5. Hit EXE, which will return the area probability (p) along with the t-scores for the lower and upper bounds.

GUIDED PRACTICE 7.5

Use a calculator to find the area to the right of $t = 3$ under the t-distribution with 35 degrees of freedom.[1]

GUIDED PRACTICE 7.6

Without doing any calculations, will the area to the right of $Z = 3$ under the standard normal curve be greater than, less than, or equal to the area to the right of $t = 3$ with 35 degrees of freedom?[2]

---

[1] Because we want to shade to the right of $t = 3$, we let lower = 3. There is no upper bound, so use a large value such as 100 for upper. Let df = 35. The area is 0.0025 or 0.25%.

[2] Because the t-distribution has greater spread and thicker tails than the normal distribution, we would expect the upper tail area to the right of $Z = 3$ to be less than the upper tail area to the right of $t = 3$. One can confirm that the area to the right of $Z = 3$ is 0.0013, which is less than 0.0025. With a smaller degrees of freedom, this difference would be even more pronounced. Try it!

7.1.4 Checking conditions for inference on a mean using the t-distribution

Using the t-distribution for inference on a mean requires two assumptions, namely that the observations are independent and that the theoretical sampling distribution of the sample mean $\bar{x}$ is nearly normal. In practice, we check whether these assumptions are reasonable by verifying that certain conditions are met.

Independent. Observations can be considered independent when the data are collected from a *random process*, such as tossing a coin, or from a *random sample*. Without a random sample or process, the standard error formula would not apply, and it is unclear to what population the inference would apply. Recall from Chapter 6 that when sampling without replacement from a finite population, the observations can be considered independent when sampling less than 10% of the population.

Nearly normal sampling distribution. We saw in Section 4.2 that the sampling distribution of a sample mean will be nearly normal when the sample is drawn from a nearly normal population or when the sample size is at least 30 ($n \geq 30$).

What should we do when the sample size is small and we are not sure whether the population distribution is nearly normal? In this case, the best we can do is look at the data for excessive skew. If the data are very skewed or have obvious outliers, this suggests that the sample did not come from a nearly normal population. However, if the data do not show obvious skew or outliers, then the idea of a nearly normal population is generally considered *reasonable*, making the assumption of a nearly normal sampling distribution for $\bar{x}$ reasonable as well.

Note that by looking at a small data set, we cannot *prove* that the population distribution is nearly normal. However, the data can suggest to us whether the population distribution being nearly normal is an unreasonable assumption.

THE NORMALITY CONDITION WITH SMALL SAMPLES

If the sample is small and there is strong skew or extreme outliers in the data, the population from which the sample was drawn may not be nearly normal.

Ideally, we use a graph of the data to check for strong skew or outliers. When the full data set is not available, summary statistics can also be used.

As the sample size goes up, it becomes less necessary to check for skew in the data. If the sample size is 30 or more, it is no longer necessary that the population distribution be nearly normal. When the sample size is large, the Central Limit Theorem tells us that the sampling distribution of the sample mean will be nearly normal regardless of the distribution of the population.

7.1.5 One sample t-interval for a mean

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who eat them.

We would like to create a confidence interval to estimate the average mercury content in dolphin muscles. We will use a sample of 19 Risso's dolphins from the Taiji area in Japan. The data are summarized in Figure 7.7.

Because we are estimating a mean, we would like to construct a t-interval, but first we must check whether the conditions for using a t-interval are met. We will start by assuming that the sample of 19 Risso's dolphins constitutes a random sample. Next, we note that the sample size is small (less than 30), and we do not know whether the distribution of mercury content for all dolphins is nearly normal. Therefore, we must look at the data. Since we do not have all of the data to graph, we look at the summary statistics provided in Figure 7.7. These summary statistics do not suggest any strong skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, we believe it is reasonable that the population distribution of mercury content in dolphins could be nearly normal.

Figure 7.6: A Risso's dolphin.

Photo by Mike Baird (www.bairdphotos.com). CC BY 2.0 license.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Figure 7.7: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

With both conditions met, we will construct a 95% confidence interval. Recall that a confidence interval has the following form:

$$\text{point estimate } \pm \text{ critical value} \times SE \text{ of estimate}$$

The point estimate is the sample mean and the SE of the sample mean is given by $s/\sqrt{n}$. What do we use for the critical value? Since we are using the t-distribution, we use a t-table to find the critical value. We denote the critical value $t^\star$.

- For a 95% confidence interval, we want to find the cutoff $t^\star$ such that 95% of the t-distribution is between -$t^\star$ and $t^\star$.

- Using the t-table on page 373, we look at the row that corresponds to the degrees of freedom and the column that corresponds to the confidence level.

DEGREES OF FREEDOM FOR A SINGLE SAMPLE

If the sample has n observations and we are examining a single mean, then we use the t-distribution with $df = n - 1$ degrees of freedom.

EXAMPLE 7.7

Calculate a 95% confidence interval for the average mercury content in dolphin muscles based on this sample. Recall that $n = 19$, $\bar{x} = 4.4$ μg/wet g, and $s = 2.3$ μg/wet g.

To find the critical value $t^\star$ we use the t-distribution with $n - 1$ degrees of freedom. The sample size is 19, so $df = 19 - 1 = 18$ degrees of freedom. Using the t-table with row $df = 18$ and column corresponding to a 95% confidence level, we get $t^\star = 2.10$. The point estimate is the sample mean $\bar{x}$ and the standard error of a sample mean is given by $\frac{s}{\sqrt{n}}$. Now we have all the pieces we need to calculate a 95% confidence interval for the average mercury content in dolphin muscles.

$$\text{point estimate } \pm \text{ critical value} \times SE \text{ of estimate}$$

$$\bar{x} \ \pm \ t^\star \times \frac{s}{\sqrt{n}} \quad df = n - 1$$

$$4.4 \ \pm \ 2.10 \times \frac{2.3}{\sqrt{19}} \quad df = 18$$

$$= (3.29,\ 5.51)$$

EXAMPLE 7.8

How do we interpret this 95% confidence interval? To what population is it applicable?

A random sample of Risso's dolphins was taken from the Taiji area in Japan. The mercury content in the muscles of other types of dolphins and from dolphins from other regions may vary. Therefore, we can only make an inference to Risso's dolphins from this area. We are 95% confident the true average mercury content in the muscles of Risso's dolphins in the Taiji area of Japan is between 3.29 and 5.51 μg/wet gram.

CONSTRUCTING A CONFIDENCE INTERVAL FOR A MEAN

To carry out a complete confidence interval procedure to estimate a single mean μ,

Identify: Identify the parameter and the confidence level, C%.

The parameter will be an unknown population mean, e.g. the true mean (or average) mercury content in Risso's dolphins.

Choose: Choose the appropriate interval procedure and identify it by name.

Here we choose the **1-sample *t*-interval**.

Check: Check conditions for the sampling distribution of $\bar{x}$ to nearly normal.

1. Data come from a random sample or random process.
2. The sample size $n \geq 30$ or the population distribution is nearly normal.

   If the sample size is less than 30 and the population distribution is unknown, check for strong skew or outliers in the data. If neither is found, the condition that the population distribution is nearly normal is considered reasonable.

Calculate: Calculate the confidence interval and record it in interval form.

point estimate $\pm\ t^\star \times SE$ of estimate, $\quad df = n - 1$

point estimate: the sample mean $\bar{x}$

SE of estimate: $\frac{s}{\sqrt{n}}$

$t^\star$: use a *t*-table at row $df = n - 1$ and confidence level C%

(___, ___)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

Here, we are C% confident that the true *mean* of [...] is between ___ and ___. A conclusion depends upon whether the interval is entirely above, is entirely below, or contains the value of interest.

EXAMPLE 7.9

The FDA's webpage provides some data on mercury content of fish.[3]   Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. Construct an appropriate 95% confidence interval for the true average mercury content of croaker white fish (Pacific). Is there evidence that the average mercury content is greater than 0.275 ppm? Use the five step framework to organize your work.

———————

Identify: The parameter of interest is the true mean mercury content in croaker white fish (Pacific). We want to estimate this at the 95% confidence level.

Choose: Because the parameter to be estimated is a single mean, we will use a 1-sample t-interval.

Check: We will assume that the sample constitutes a random sample of croaker white fish (Pacific). The sample size n is small, but there are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not too great. Therefore we think it is reasonable that the population distribution of mercury content in croaker white fish (Pacific) could be nearly normal.

Calculate: We will calculate the interval:

$$\text{point estimate } \pm \ t^\star \times SE \text{ of estimate}$$

The point estimate is the sample mean: $\bar{x} = 0.287$

The SE of the sample mean is: $\frac{s}{\sqrt{n}} = \frac{0.069}{\sqrt{15}}$

We find $t^\star$ for the one sample case using the t-table at row $df = n - 1$ and confidence level C%. For a 95% confidence level and $df = 15 - 1 = 14$, $t^\star = 2.145$.

So the 95% confidence interval is given by:

$$0.287 \pm 2.145 \times \frac{0.069}{\sqrt{15}} \qquad df = 14$$
$$0.287 \pm 2.145 \times 0.0178$$
$$= (0.249, \ 0.325)$$

Conclude: We are 95% confident that the true *average* mercury content of croaker white fish (Pacific) is between 0.249 and 0.325 ppm. Because the interval contains 0.275 as well as values less than 0.275, we do not have evidence that the true average mercury content is greater than 0.275 ppm.

EXAMPLE 7.10

Based on the interval calculated in Example 7.9 above, can we say that 95% of croaker white fish (Pacific) have mercury content between 0.249 and 0.325 ppm?

———————

No. The interval estimates the *average* amount of mercury with 95% confidence. It is not trying to capture 95% of the values.

———————————————

[3]www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

7.1.6 Calculator: the 1-sample *t*-interval

TI-83/84: 1-SAMPLE T-INTERVAL

Use STAT, TESTS, TInterval.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose 8:TInterval.
4. Choose Data if you have all the data or Stats if you have the mean and standard deviation.

   - If you choose Data, let List be L1 or the list in which you entered your data (don't forget to enter the data!) and let Freq be 1.
   - If you choose Stats, enter the mean, SD, and sample size.

5. Let C-Level be the desired confidence level.
6. Choose Calculate and hit ENTER, which returns:

(__,__)	the confidence interval
$\bar{x}$	the sample mean
Sx	the sample SD
n	the sample size

CASIO FX-9750GII: 1-SAMPLE T-INTERVAL

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. If necessary, enter the data into a list.
3. Choose the INTR option (F3 button), t (F2 button), and 1-S (F1 button).
4. Choose either the Var option (F2) or enter the data in using the List option.
5. Specify the interval details:

   - Confidence level of interest for C-Level.
   - If using the Var option, enter the summary statistics. If using List, specify the list and leave Freq value at 1.

6. Hit the EXE button, which returns

Left, Right	ends of the confidence interval
$\bar{x}$	sample mean
sx	sample standard deviation
n	sample size

GUIDED PRACTICE 7.11

Use a calculator to find a 95% confidence interval for the mean mercury content in croaker white fish (Pacific). The sample size was 15, and the sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively.[4]

---

[4]Choose TInterval or equivalent. We do not have all the data, so choose Stats on a TI or Var on a Casio. Enter $\bar{x}$ and Sx. Note: Sx is the sample standard deviation (0.069), not the SE. Let n = 15 and C-Level = 0.95. This should give the interval (0.249, 0.325).

7.1.7   Choosing a sample size when estimating a mean

In Section 6.1.5, we looked at sample size considerations when estimating a proportion. We take the same approach when estimating a mean. Recall that the margin of error is measured as the distance between the point estimate and the upper or lower bound of the confidence interval. We want to estimate a mean with a particular confidence level while putting an upper bound on the margin of error. What is the smallest sample size that will satisfy these conditions?

For a one sample t-interval, the margin of error, ME, is given by $ME = t^\star \times \frac{s}{\sqrt{n}}$. The challenge in this case is that we need to know n to find $t^\star$. But n is precisely what we are attempting to solve for! Fortunately, in most cases we will have a reasonable estimate for the population standard deviation and the desired n will be large, so we can use $ME = z^\star \times \frac{\sigma}{\sqrt{n}}$, making it easier to solve for n.

EXAMPLE 7.12

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The standard deviation of systolic blood pressure for people in the U.S. is about 25 mmHg (millimeters of mercury). How large of a sample is necessary to estimate the average systolic blood pressure of people in a particular town with a margin of error no greater than 4 mmHg using a 95% confidence level?

―――――――

For this problem, we want to find the sample size n so that the margin of error, ME, is less than or equal to 4 mmHg. We start by writing the following inequality:

$$z^\star \times \frac{\sigma}{\sqrt{n}} \le 4$$

For a 95% confidence level, the critical value $z^\star = 1.96$. Our best estimate for the population standard deviation is $\sigma = 25$. We substitute in these two values and we solve for n.

$$1.96 \times \frac{25}{\sqrt{n}} \le 4$$
$$1.96 \times \frac{25}{4} \le \sqrt{n}$$
$$\left(1.96 \times \frac{25}{4}\right)^2 \le n$$
$$150.06 \le n$$
$$n = 151$$

The minimum sample size that meets the condition is 151. We round up because the sample size must be an integer and it must be *greater than or equal to* 150.06.

IDENTIFY A SAMPLE SIZE FOR A PARTICULAR MARGIN OF ERROR

To estimate the minimum sample size required to achieve a margin of error less than or equal to m, with C% confidence, we set up an inequality as follows:

$$z^\star \frac{\sigma}{\sqrt{n}} \le m$$

$z^\star$ depends on the desired confidence level and σ is the standard deviation associated with the population. We solve for the sample size, n.

Sample size computations are helpful in planning data collection, and they require careful forethought.

7.1.8 Hypothesis testing for a mean

Is the typical U.S. runner getting faster or slower over time? Technological advances in shoes, training, and diet might suggest runners would be faster. An opposing viewpoint might say that with the average body mass index on the rise, people tend to run slower. In fact, all of these components might be influencing run time.

We consider this question in the context of the Cherry Blossom Race, which is a 10-mile race in Washington, DC each spring. The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.3 minutes (93 minutes and about 18 seconds). We want to determine using data from 100 participants in the 2017 Cherry Blossom Race whether runners in this race are getting faster or slower, versus the other possibility that there has been no change. Figure 7.8 shows run times for 100 randomly selected participants.
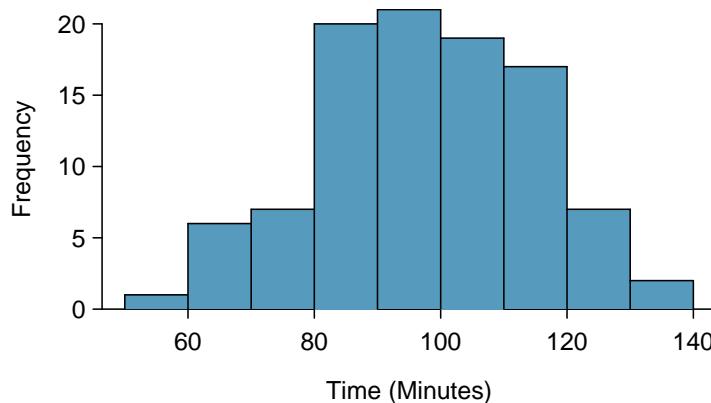
Figure 7.8: A histogram of `time` for the sample of 2017 Cherry Blossom Race participants.

EXAMPLE 7.13

What are appropriate hypotheses for this context?

We know that the average run time for all runners in 2006 was 93.3 minutes. We have a sample of times from the 2017 race. We are interested in whether the average run time has *changed*, so we will use a two-sided H_A.

Let μ represent the average 10-mile run time of all participants in 2017, which is unknown to us.

H_0: $\mu = 93.3$ minutes. The average run time of all participants in 2017 was 93.3 min.
H_A: $\mu \neq 93.3$ minutes. The average run time of all participants in 2017 was not 93.3 min.

The data come from a random sample from a large population, so the observations are independent. Do we need to check for skew in the data? No – with a sample size of 100, well over 30, the Central Limit Theorem tells us that the sampling distribution of $\bar{x}$ will be nearly normal.

With independence satisfied and slight skew not a concern for this large of a sample, we can proceed with performing a hypothesis test using the t-distribution.

The sample mean and sample standard deviation of the 100 runners from the 2017 Cherry Blossom Race are 97.3 and 17.0 minutes, respectively. We want to know whether the observed sample mean of 97.3 is far enough away from 93.3 to provide convincing evidence of a real difference, or if it is within the realm of expected variation for a sample of size 100.

To answer this question we will find the test statistic and p-value for the hypothesis test. Since we will be using a sample standard deviation in our calculation of the test statistic, we will need to use a t-distribution, just as we did with confidence intervals for a mean. We call the test statistic a T-statistic. It has the same general form as a Z-statistic.

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

As we saw before, when carrying out inference on a single mean, the degrees of freedom is given by $n - 1$.

THE T-STATISTIC

The **T-statistic** (or T-score) is analogous to a Z-statistic (or Z-score). Both represent how many standard errors the observed value is from the null value.

EXAMPLE 7.14

Calculate the test statistic, degrees of freedom, and p-value for this test.

———————

Here, our point estimate is the sample mean, $\bar{x} = 97.3$ minutes.

The SE of the sample mean is given by $\frac{s}{\sqrt{n}}$, so the SE of estimate $= \frac{17.0}{\sqrt{100}} = 1.7$ minutes.

$$T = \frac{97.3 - 93.3}{1.7} = 2.35 \qquad df = 100 - 1 = 99$$

Using a calculator, we find that the area above 2.35 under the t-distribution with 99 degrees of freedom is 0.01. Because this is a two-tailed test, we double this. So the p-value $= 2 \times 0.01 = 0.02$.

EXAMPLE 7.15

Does the data provide sufficient evidence that the average Cherry Blossom Run time in 2017 is different than in 2006?

———————

This depends upon the desired significance level. Since the p-value $= 0.02 < 0.05$, there is sufficient evidence at the 5% significance level. However, as the p-value of $0.02 > 0.01$, there is not sufficient evidence at the 1% significance level.

EXAMPLE 7.16

Would you expect the hypothesized value of 93.3 to fall inside or outside of a 95% confidence interval? What about a 99% confidence interval?

———————

Because the hypothesized value of 93.3 was rejected by the two-sided $\alpha = 0.05$ test, we would expect it to be outside the 95% confidence interval. However, because the hypothesized value of 93.3 was not rejected by the two-sided $\alpha = 0.01$ test, we would expect it to fall inside the (wider) 99% confidence interval.

HYPOTHESIS TEST FOR A MEAN

To carry out a complete hypothesis test to test the claim that a single mean μ is equal to a null value μ_0,

Identify: Identify the hypotheses and the significance level, α.

H_0: $\mu = \mu_0$

H_A: $\mu \neq \mu_0$; H_A: $\mu > \mu_0$;   or H_A: $\mu < \mu_0$

Choose: Choose the appropriate test procedure and identify it by name.

Here we choose the **1-sample t-test**.

Check: Check conditions for the sampling distribution of $\bar{x}$ to be nearly normal.

1. Data come from a random sample or random process.

2. The sample size $n \geq 30$ or the population distribution is nearly normal.

   If the sample size is less than 30 and the population distribution is unknown, check for strong skew or outliers in the data. If neither is found, then the condition that the population is nearly normal is considered reasonable.

Calculate: Calculate the t-statistic, df, and p-value.

$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}, \quad df = n - 1$

point estimate: the sample mean $\bar{x}$

SE of estimate: $\frac{s}{\sqrt{n}}$

null value: μ_0

p-value = (based on the t-statistic, the df, and the direction of H_A)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].

EXAMPLE 7.17

Recall the example involving the mercury content in croaker white fish (Pacific). Based on a sample of size 15, a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. Carry out an appropriate test to determine if 0.25 is a reasonable value for the average mercury content of croaker white fish (Pacific). Use the five step method to organize your work.

Identify: We will test the following hypotheses at the $\alpha = 0.05$ significance level.
  H_0: $\mu = 0.25$
  H_A: $\mu \neq 0.25$ The mean mercury content is not 0.25 ppm.

Choose: Because we are hypothesizing about a single mean we choose the 1-sample t-test.

Check: The conditions were checked previously, namely – the data come from a random sample, and because n is less than 30, we verified that there is no strong skew or outliers in the data, so the assumption that the population distribution of mercury is nearly normally distributed is reasonable.

Calculate: We will calculate the t-statistic and the p-value.

$$T = \frac{\text{point estimate } - \text{ null value}}{SE \text{ of estimate}}$$

The point estimate is the sample mean: $\bar{x} = 0.287$

The SE of the sample mean is: $\frac{s}{\sqrt{n}} = \frac{0.069}{\sqrt{15}} = 0.0178$

The null value is the value hypothesized for the parameter in H_0, which is 0.25.

For the 1-sample t-test, $df = n - 1$.

$$T = \frac{0.287 - 0.25}{0.0178} = 2.07 \qquad df = 15 - 1 = 14$$

Because H_A is a two-tailed test ($\neq$), the p-value corresponds to the area to the right of $t = 2.07$ plus the area to the left of $t = -2.07$ under the t-distribution with 14 degrees of freedom. The p-value $= 2 \times 0.029 = 0.058$.

Conclude: The p-value of $0.058 > 0.05$, so we do not reject the null hypothesis. We do not have sufficient evidence that the average mercury content in croaker white fish (Pacific) is not 0.25.

GUIDED PRACTICE 7.18

Recall that the 95% confidence interval for the average mercury content in croaker white fish was (0.249, 0.325). Discuss whether the conclusion of the hypothesis test in the previous example is consistent or inconsistent with the conclusion of the confidence interval.[5]

---

[5]It is consistent because 0.25 is located (just barely) inside the confidence interval, so it is considered a reasonable value. Our hypothesis test did not reject the hypothesis that $\mu = 0.25$, also implying that it is a reasonable value. Note that the p-value was just over the cutoff of 0.05. This is consistent with the value of 0.25 being just inside the confidence interval. Also note that the hypothesis test did not *prove* that $\mu = 0.25$. The value 0.25 is just one of many reasonable values for the true mean.

7.1.9 Calculator: 1-sample *t*-test

TI-83/84: 1-SAMPLE T-TEST

Use STAT, TESTS, T-Test.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose 2:T-Test.
4. Choose Data if you have all the data or Stats if you have the mean and standard deviation.
5. Let μ_0 be the null or hypothesized value of μ.

   - If you choose Data, let List be L1 or the list in which you entered your data (don't forget to enter the data!) and let Freq be 1.
   - If you choose Stats, enter the mean, SD, and sample size.

6. Choose $\neq$, $<$, or $>$ to correspond to H_A.
7. Choose Calculate and hit ENTER, which returns:

t	t statistic	Sx	the sample standard deviation
p	p-value	n	the sample size
x̄	the sample mean		

CASIO FX-9750GII: 1-SAMPLE T-TEST

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. If necessary, enter the data into a list.
3. Choose the TEST option (F3 button).
4. Choose the t option (F2 button).
5. Choose the 1-S option (F1 button).
6. Choose either the Var option (F2) or enter the data in using the List option.
7. Specify the test details:

   - Specify the sidedness of the test using the F1, F2, and F3 keys.
   - Enter the null value, $\mu0$.
   - If using the Var option, enter the summary statistics. If using List, specify the list and leave Freq values at 1.

8. Hit the EXE button, which returns

	alternative hypothesis	x̄	sample mean
t	T statistic	sx	sample standard deviation
p	p-value	n	sample size

GUIDED PRACTICE 7.19

The average time for all runners who finished the Cherry Blossom Run in 2006 was 93.3 minutes. In 2017, the average time for 100 randomly selected participants was 97.3, with a standard deviation of 17.0 minutes. Use a calculator to find the T-statistic and p-value for the appropriate test to see if the average time for the participants in 2017 is different than it was in 2006.[6]

---

[6]Choose T-Test or equivalent. Let μ_0 be 93.3. x̄ is 97.3, S_x is 17.0, and n = 100. Choose $\neq$ to correspond to H_A. We get t = 2.353 and the p-value p = 0.021.

Section summary

- The t-distribution.

  - When calculating a test statistic for a mean, using the sample standard deviation in place of the population standard deviation gives rise to a new distribution called the t-distribution.
  - As the sample size and degrees of freedom increase, s becomes a more stable estimate of σ, and the corresponding t-distribution has smaller spread.
  - As the degrees of freedom go to ∞, the t-distribution approaches the normal distribution. This is why we can use the t-table at $df = \infty$ to find the value of $z^\star$.

- When carrying out inference for a single mean, we use the t-distribution with $n - 1$ degrees of freedom.

- When there is one sample and the parameter of interest is a single mean:

  - Estimate μ at the C% confidence level using a **1-sample t-interval**.
  - Test H_0: $\mu = \mu_0$ at the α significance level using a **1-sample t-test**.

- The conditions for the one sample t-interval and t-test are the same.

  1. The data come from a random sample or random process.
  2. The sample size $n \geq 30$ or the population distribution is nearly normal.

     If the sample size is less than 30 and the population distribution is unknown, check for strong skew or outliers in the data. If neither is found, then the condition that the population distribution is nearly normal is considered reasonable.

- When the conditions are met, we calculate the confidence interval and the test statistic as we did in the previous chapter, except that we use $t^\star$ for the critical value and we use T for the test statistic.

  Confidence interval: point estimate $\pm\ t^\star \times SE$ of estimate

  Test statistic: $T = \frac{\text{point estimate } - \text{ null value}}{SE \text{ of estimate}}$

  Here the point estimate is the sample mean: $\bar{x}$.

  The SE of estimate is the SE of the sample mean: $\frac{s}{\sqrt{n}}$.

  The degrees of freedom is given by $df = n - 1$.

- To calculate the minimum sample size required to estimate a mean with C% confidence and a margin of error no greater than m, we set up an inequality as follows:
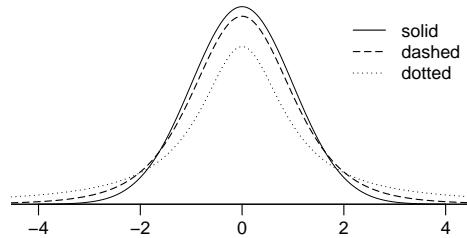
$$z^\star \frac{\sigma}{\sqrt{n}} \leq m$$

  $z^\star$ depends on the desired confidence level and σ is the standard deviation associated with the population. We solve for the sample size, n. Always round the answer up to the next *integer*, since n refers to a number of people or things.

Exercises

7.1 Identify the critical t. 📹 An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t-value ($t^\star$) for the given sample size and confidence level.

(a) $n = 6$, CL $= 90\%$

(b) $n = 21$, CL $= 98\%$

(c) $n = 29$, CL $= 95\%$

(d) $n = 12$, CL $= 99\%$

7.2 t-distribution. The figure on the right shows three unimodal and symmetric curves: the standard normal (z) distribution, the t-distribution with 5 degrees of freedom, and the t-distribution with 1 degree of freedom. Determine which is which, and explain your reasoning.

7.3 Find the p-value, Part I. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size and test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

(a) $n = 11$, $T = 1.91$

(b) $n = 17$, $T = -3.45$

(c) $n = 7$, $T = 0.83$

(d) $n = 28$, $T = 2.13$

7.4 Find the p-value, Part II. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size and test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.01$.

(a) $n = 26$, $T = 2.485$

(b) $n = 18$, $T = 0.5$

7.5 Working backwards, Part I. A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). This confidence interval is based on a simple random sample of 36 observations. Calculate the sample mean and standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t-distribution in any calculations.
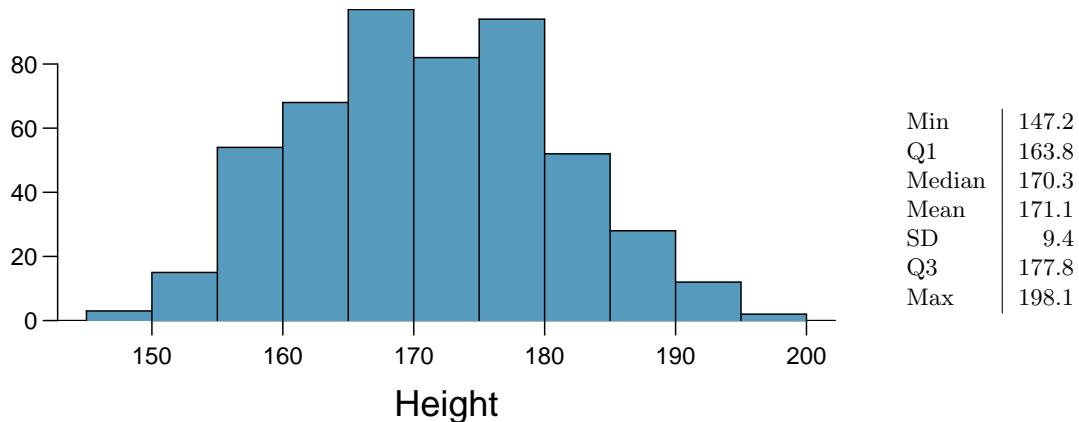
7.6 Working backwards, Part II. A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

7.7 Sleep habits of New Yorkers. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Is the result statistically significant?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(a) Write the hypotheses in symbols and in words.

(b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

(c) Find and interpret the p-value in this context. Drawing a picture may be helpful.

(d) What is the conclusion of the hypothesis test?

(e) If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

7.8 Heights of adults. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.[7]

Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

(a) What is the point estimate for the average height of active individuals? What about the median?

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

7.9 Find the mean. You are given the following hypotheses:

$$H_0 : \mu = 60$$
$$H_A : \mu \neq 60$$

We know that the sample standard deviation is 8 and the sample size is 20. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

---

[7]G. Heinz et al. "Exploring relationships in body dimensions". In: *Journal of Statistics Education* 11.2 (2003).

7.10 $t^\star$ **vs.** $z^\star$**.** For a given confidence level, $t^\star_{df}$ is larger than $z^\star$. Explain how $t^\star_{df}$ being slightly larger than $z^\star$ affects the width of the confidence interval.

7.11 Play the piano. Georgianna claims that in a small city renowned for its music school, the average child takes at least 5 years of piano lessons. We have a random sample of 30 children from the city, with a mean of 4.6 years of piano lessons and a standard deviation of 2.2 years.

(a) Use a hypothesis test to determine if there is sufficient evidence against Georgianna's claim.

(b) Construct a 95% confidence interval for the number of years students in this city take piano lessons, and interpret it in context of the data.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

7.12 Auto exhaust and lead exposure. Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 μg/l and a SD of 37.74 μg/l; a previous study of individuals from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 μg/l.[8]

(a) Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a different concentration of lead.

(b) Explicitly state and check all conditions necessary for inference on these data.

(c) Regardless of your answers in part (b), test the hypothesis that the downtown police officers have a higher lead exposure than the group in the previous study. Interpret your results in context.

7.13 Car insurance savings. A market researcher wants to evaluate car insurance savings at a competing company. Based on past studies he is assuming that the standard deviation of savings is $100. He wants to collect data such that he can get a margin of error of no more than $10 at a 95% confidence level. How large of a sample should he collect?

7.14 SAT scores. The standard deviation of SAT scores for students at a particular Ivy League college is 250 points. Two statistics students, Raina and Luke, want to estimate the average SAT score of students at this college as part of a class project. They want their margin of error to be no more than 25 points.

(a) Raina wants to use a 90% confidence interval. How large a sample should she collect?

(b) Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina's, and explain your reasoning.

(c) Calculate the minimum required sample size for Luke.

---

[8]WI Mortada et al. "Study of lead exposure from automobile exhaust as a risk for nephrotoxicity among traffic policemen." In: *American journal of nephrology* 21.4 (2000), pp. 274–279.

7.2 Inference for paired data

When we have two observations on each person or each case, we can answer questions such as the following:

- Do students do better on reading or writing sections of standardized tests?

- How do the number of days with temperature above 90°F compare between 1948 and 2018?

- Are Amazon textbook prices lower than the college bookstore prices? If so, how much lower, on average?

Learning objectives

1. Distinguish between paired and unpaired data.

2. Recognize that inference procedures for paired data use the same one-sample t-procedures as in the previous section, and that these procedures are applied to the *differences* of the paired observations.

3. Carry out a complete hypothesis test for paired differences.

4. Carry out a complete confidence interval procedure for paired differences.

7.2.1 Paired observations and samples

In the previous edition of this textbook, we found that Amazon prices were, on average, lower than those of the UCLA Bookstore for UCLA courses in 2010. It's been several years, and many stores have adapted to the online market, so we wondered, how is the UCLA Bookstore doing today?

We sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.9, where prices are in U.S. dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
67	Korean	1	24.96	23.79	1.17
68	Jewish Studies	M10	35.96	32.40	3.56

Figure 7.9: Five cases of the `textbooks` data set.

Each textbook has two corresponding prices in the data set: one for the UCLA Bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.
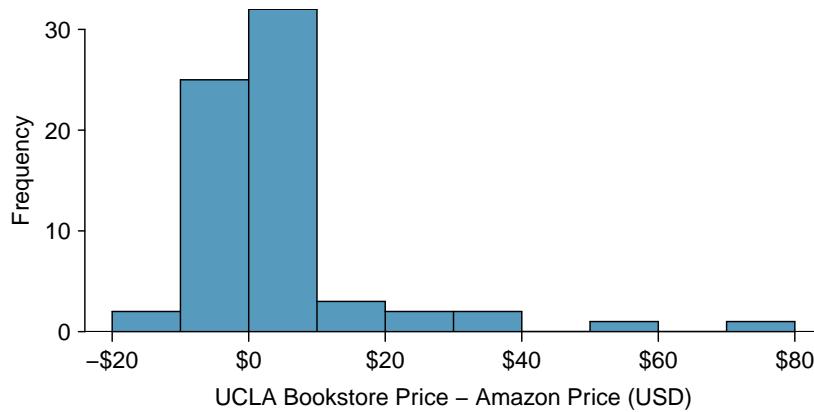
Figure 7.10: Histogram of the difference in price for each book sampled. These data are very strongly skewed. . Explore this data set on Tableau Public⁺.

PAIRED DATA

Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the `textbook` data set, we look at the differences in prices, which is represented as the `diff` variable. Here, for each book, the differences are taken as

$$\text{UCLA Bookstore price} - \text{Amazon price}$$

It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure 7.10. Using differences between paired observations is a common and useful way to analyze paired data.

GUIDED PRACTICE 7.20

The first difference shown in Figure 7.9 is computed as: $47.97 - 47.45 = 0.52$. What does this difference tell us about the price for this textbook on Amazon versus the UCLA bookstore?[9]

7.2.2 Hypothesis tests for paired data

To analyze a paired data set, we simply analyze the differences. We can use the same t-distribution techniques we applied in the last section.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

Figure 7.11: Summary statistics for the price differences. There were 68 books, so there are 68 differences.

---

[9]The difference is taken as UCLA Bookstore price $-$ Amazon price. Because the difference is positive, it tells us that the UCLA Bookstore price was *greater* for this textbook. In fact, it was \$0.52, or 52 cents, more expensive at the UCLA bookstore than on Amazon.

We will set up and implement a hypothesis test to determine whether, on average, there is a difference in textbook prices between Amazon and the UCLA bookstore. We are considering two scenarios: there is no difference in prices or there is some difference in prices.

H_0: $\mu_{diff} = 0$. On average, there is no difference in textbook prices.

H_A: $\mu_{diff} \neq 0$. On average, there is some difference in textbook prices.

Can the t-distribution be used for this application? The observations are based on a random sample from a large population, so independence is reasonable. While the distribution of the data is very strongly skewed, we do have $n = 68$ observations. This sample size is large enough that we do not have to worry about whether the population distribution for difference in price might be nearly normal or not. Because the conditions are satisfied, we can use the t-distribution to this setting.

We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 13.42$) and the number of differences ($n_{diff} = 68$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$$

Next we compute the test statistic. The point estimate is the observed value of $\bar{x}_{diff}$. The null value is the value hypothesized under the null hypothesis. Here, the null hypothesis is that the true mean of the differences is 0.

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}} = \frac{3.58 - 0}{1.63} = 2.20$$

The degrees of freedom are $df = 68 - 1 = 67$. To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true. This is shown in Figure 7.12. Because this is a two-sided test, the p-value corresponds to the area in both tails. Using statistical software, we find the area in the tails to be 0.0312.

Because the p-value of 0.0312 is less than 0.05, we reject the null hypothesis. We have evidence that, on average, there is a difference in textbook prices. In particular, we can say that, on average, Amazon prices are lower than the UCLA Bookstore prices for UCLA course textbooks.

$\mu_0 = 0$ $\bar{x}_{diff} = 2.98$
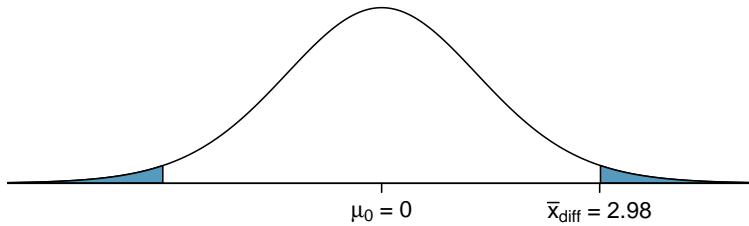
Figure 7.12: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

EXAMPLE 7.21

We have evidence to conclude Amazon is, on average, less expensive. Does this mean that UCLA students should always buy their books on Amazon?

No. The fact that Amazon is, on average, less expensive, does not imply that it is less expensive for *every* textbook. Examining the distribution shown in Figure 7.10, we see that there are certainly a handful of cases where Amazon prices are much lower than the UCLA Bookstore's, which suggests it is worth checking Amazon or other online sites before purchasing. However, in many cases the Amazon price is *above* what the UCLA Bookstore charges, and most of the time the price isn't that different.

For reference, this is a very different result from what we (the authors) had seen in a similar data set from 2010. At that time, Amazon prices were almost uniformly lower than those of the UCLA Bookstore's and by a large margin, making the case to use Amazon over the UCLA Bookstore quite compelling at that time.

HYPOTHESIS TEST FOR PAIRED DATA

To carry out a complete hypothesis test to test the claim that a mean of differences μ_{diff} is equal to 0,

Identify: Identify the hypotheses and the significance level, α.

> H_0: $\mu_{diff} = 0$
>
> H_A: $\mu_{diff} \neq 0$; H_A: $\mu_{diff} > 0$;   or H_A: $\mu_{diff} < 0$

Choose: Choose the appropriate test procedure and identify it by name.

> Here we choose the **matched pairs t-test**.

Check: Check conditions for the sampling distribution of $\bar{x}_{diff}$ to be nearly normal.

> 1. There is paired data from a random sample or from a matched pairs experiment.
> 2. $n_{diff} \geq 30$ or the population of differences is nearly normal.
>   If the number of differences is less than 30 and the distribution of the population of differences is unknown, check for strong skew or outliers in the sample differences. If neither is found, then the condition that the population of differences is nearly normal is considered reasonable.

Calculate: Calculate the t-statistic, df, and p-value.

> $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, $df = n_{diff} - 1$
>
> > point estimate: the sample mean of differences $\bar{x}_{diff}$
> >
> > SE of estimate: $\frac{s_{diff}}{\sqrt{n_{diff}}}$
> >
> > null value: 0
>
> p-value = (based on the t-statistic, the df, and the direction of H_A)

Conclude: Compare the p-value to α, and draw a conclusion in context.

> If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].
>
> If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].
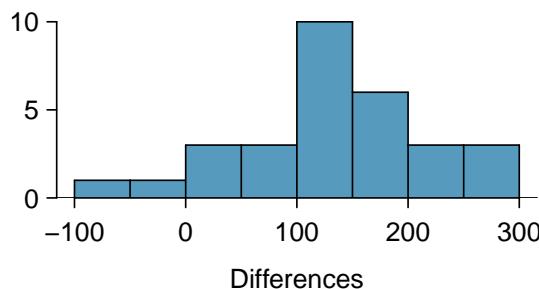
Figure 7.13: SAT score after course - SAT score before course.

EXAMPLE 7.22

An SAT preparation company claims that its students' scores improve by over 100 points on average after their course. A consumer group would like to evaluate this claim, and they collect data on a random sample of 30 students who took the class. Each of these students took the SAT before and after taking the company's course, so we have a difference in scores for each student. We will examine these differences $x_1 = 57$, $x_2 = 133$, ..., $x_{30} = 140$ as a sample to evaluate the company's claim. The distribution of the differences, shown in Figure 7.13, has a mean of 135.9 and a standard deviation of 82.2. Do these data provide convincing evidence to back up the company's claim? Use the five step framework to organize your work.

――――――――

Identify: We will test the following hypotheses at the $\alpha = 0.05$ level:
   H_0: $\mu_{diff} = 100$. Student scores improve by 100 points, on average.
   H_A: $\mu_{diff} > 100$. Student scores improve by more than 100 points, on average.

   Here, $diff$ = SAT score after course - SAT score before course.

Choose: Because we have paired data and the parameter to be estimated is a mean of differences, we will use a matched pairs t-test.

Check: We have a random sample with paired observations and the number of differences is $n_{diff} = 30 \geq 30$, so we can proceed with the matched pairs t-test.

Calculate: We will calculate the test statistic, df, and p-value.

$$T = \frac{\text{point estimate} \ - \ \text{null value}}{SE \text{ of estimate}}$$

The point estimate is the sample mean of differences: $\bar{x}_{diff} = 135.9$

The SE of the sample mean of differences is: $\frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{82.2}{\sqrt{30}} = 15.0$

Since this is essentially a one sample t-test, the degrees of freedom is $n_{diff} - 1$.

$$T = \frac{135.9 - 100}{\frac{82.2}{\sqrt{30}}} = \frac{135.9 - 100}{15.0} = 2.4 \qquad df = 30 - 1 = 29$$

The p-value is the area to the right of 2.4 under the t-distribution with 29 degrees of freedom. The p-value = 0.012.

Conclude: p-value = $0.012 < \alpha$ so we reject the null hypothesis. The data provide convincing evidence to support the company's claim that students' scores improve by more than 100 points, on average, following the class.

GUIDED PRACTICE 7.23

Because we found evidence to support the company's claim, does this mean that a student will score more than 100 points higher on the SAT if they take the class than if they do not take the class?

――――――――――――――――――

[9]No. First, this is an observational study, so we cannot make a causal conclusion. Maybe SAT test takers tend to improve their score over time even if they don't take this SAT class. Secondly, students' scores improved by more than 100 points *on average*. That does not imply that each student improved by more than 100 points. With a sample standard deviation of 82.2 and a mean of 135.9, some students did worse after the SAT class. This can be verified by Figure 7.13.

7.2.3 Calculator: the matched pairs t-test

The matched pairs t-test is a one sample t-test. Instead of using the data or the summary statistics from a sample, make sure to use the data of *differences* or the summary statistics for the *differences*.

TI-83/84: MATCHED PAIRS T-TEST

Use `STAT`, `TESTS`, `T-Test`.

1. Choose `STAT`.

2. Right arrow to `TESTS`.

3. Down arrow and choose `2:T-Test`.

4. Choose `Data` if you have all the data or `Stats` if you have the mean and standard deviation.

5. Let μ_0 be the null or hypothesized value of μ_{diff}.
   - If you choose `Data`, let `List` be `L3` or the list in which you entered the differences, and let `Freq` be `1`.
   - If you choose `Stats`, enter the mean, SD, and sample size of the differences.

6. Choose $\neq$, `<`, or `>` to correspond to H_A.

7. Choose `Calculate` and hit `ENTER`, which returns:
t	t statistic
p	p-value
x̄	the sample mean of the differences
Sx	the sample SD of the differences
n	the sample size of the differences

CASIO FX-9750GII: MATCHED PAIRS T-TEST

1. Compute the paired differences of the observations.

2. Using the computed differences, follow the instructions for a 1-sample t-test

---

7.2.4 Confidence intervals for the mean of a difference

In the previous examples, we carried out a matched pairs t-test, where the null hypothesis was that the true average of the paired differences is zero. Sometimes we want to estimate the true average of paired differences with a confidence interval, and we use a **matched pairs t-interval**. Consider again the table summarizing data on: UCLA Bookstore price − Amazon price, for each of the 68 books sampled.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

Figure 7.14: Summary statistics for the price differences. There were 68 books, so there are 68 differences.

We construct a 95% confidence interval for the average price difference between books at the UCLA Bookstore and on Amazon. Conditions have already verified, namely, that we have paired data from a random sample and that the number of differences is at least 30. We must find the critical value, $t^\star$. Since $df = 67$ is not on the t-table, round the df down to 60 to get a $t^\star$ of 2.00 for 95% confidence. (See Section 7.2.5 for how to get a more precise interval using a calculator.)

Plugging the $t^\star$ value, point estimate, and standard error into the confidence interval formula, we get:

$$\text{point estimate} \ \pm\ t^\star \times SE \text{ of estimate} \quad \rightarrow \quad 3.58 \quad \pm\ 2.00 \times \frac{13.42}{\sqrt{68}} \quad \rightarrow \quad (0.33, 6.83)$$

We are 95% confident that the UCLA bookstore is, on average, between \$0.33 and \$6.83 more expensive than Amazon for UCLA course books. This interval does not contain zero, so it is consistent with the earlier hypothesis test that *rejected* the null hypothesis that the average difference was 0. Because our interval is entirely above 0, we have evidence that the true average difference is greater than zero. Unlike the hypothesis test, the confidence interval gives us a good idea of how much more expensive the UCLA bookstore might be, on average.

EXAMPLE 7.24

Based on the interval, can we say that 95% of the books cost between \$0.33 and \$6.83 more at the UCLA Bookstore than on Amazon?

No. This interval is attempting to estimate the *average* difference with 95% confidence. It is not attempting to capture 95% of the values. A quick look at Figure 7.10 shows that much less than 95% of the differences fall between \$0.32 and \$6.84.

CONSTRUCTING A CONFIDENCE INTERVAL FOR PAIRED DATA

To carry out a complete confidence interval procedure to estimate a mean of differences μ_{diff},

Identify: Identify the parameter and the confidence level, C%.

   The parameter will be a mean of differences, e.g. the true mean of the differences in county population (year $2018 -$ year 2017).

Choose: Choose the appropriate interval procedure and identify it by name.

   Here we choose the **matched pairs t-interval**.

Check: Check conditions for the sampling distribution of $\bar{x}_{diff}$ to be nearly normal.

   1. There is paired data from a random sample or from a matched pairs experiment.
   2. $n_{diff} \geq 30$ or the population of differences is nearly normal.

      If the number of differences is less than 30 and the distribution of the population

      of differences is unknown, check for strong skew or outliers in the sample differences.

      If neither is found, then the condition that the population of differences is nearly normal

      is considered reasonable.

Calculate: Calculate the confidence interval and record it in interval form.

   $$\text{point estimate} \ \pm\ t^\star \times SE \text{ of estimate}, \quad df\text{: } n_{diff} - 1$$

      point estimate: the sample mean of differences $\bar{x}_{diff}$

      SE of estimate: $\frac{s_{diff}}{\sqrt{n_{diff}}}$

      $t^\star$: use a t-table at row $df = n_{diff} - 1$ and confidence level C%

   (___, ___)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

   We are C% confident that the true *mean of the differences in* [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.
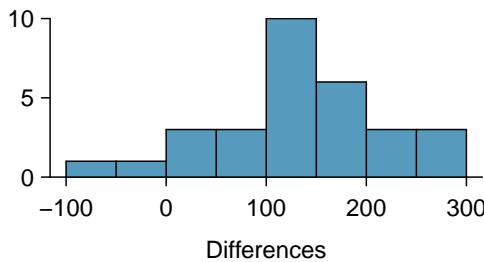
Figure 7.15: SAT score after course - SAT score before course.

EXAMPLE 7.25

An SAT preparation company claims that its students' scores improve by over 100 points on average after their course. A consumer group would like to evaluate this claim, and they collect data on a random sample of 30 students who took the class. Each of these students took the SAT before and after taking the company's course, so we have a difference in scores for each student. We will examine these differences $x_1 = 57$, $x_2 = 133$, ..., $x_{30} = 140$ as a sample to evaluate the company's claim. The distribution of the differences, shown in Figure 7.15, has a mean of 135.9 and a standard deviation of 82.2. Construct a confidence interval to estimate the true average increase in SAT after taking the company's course. Is there evidence at the 95% confidence level that students score an average of more than 100 points higher after the class? Use the five step framework to organize your work.

———————

Identify: The parameter we want to estimate is μ_{diff}, the true change in SAT score after taking the company's course. Here, $diff$ = SAT score after course $-$ SAT score before course. We will estimate this parameter at the 95% confidence level.

Choose: Because we have paired data and the parameter to be estimated is a mean of differences, we will use a matched pairs t-interval.

Check: We have a random sample with paired observations and the number of differences is $n_{diff} = 30 \geq 30$, so we can proceed with the matched pairs t-interval.

Calculate: We will calculate the confidence interval as follows.

$$\text{point estimate} \ \pm \ t^\star \times SE \text{ of estimate}$$

The point estimate is the sample mean of differences: $\bar{x}_{diff} = 135.9$

The SE of the sample mean of differences is: $\frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{82.2}{\sqrt{30}} = 15.0$

We find $t^\star$ for the one sample case using the t-table at row $df = n - 1$ and confidence level C%. For a 95% confidence level and $df = 30 - 1 = 29$, $t^\star = 2.045$.

The 95% confidence interval is given by:

$$135.9 \pm 2.045 \times \frac{82.2}{\sqrt{30}} \qquad df = 15$$
$$135.9 \pm 2.045 \times 15.0$$
$$= (105.2, \ 166.6)$$

Conclude: We are 95% confident that the true *average* increase in SAT score following the company's course is between 105.2 points to 166.6 points. There is sufficient evidence that students score greater than 100 points higher, on average, after the company's course because the entire interval is above 100.

GUIDED PRACTICE 7.26

(G)

Based on the interval (105.2, 166.6), calculated previously, can we say that 95% of student scores increased between 105.2 and 166.6 points after taking the company's course?

7.2.5   Calculator: the matched pairs t-interval

TI-83/84: MATCHED PAIRS T-INTERVAL

Use STAT, TESTS, TInterval.

1. Choose STAT.

2. Right arrow to TESTS.

3. Down arrow and choose 8:TInterval.

4. Choose Data if you have all the data or Stats if you have the mean and standard deviation.

   - If you choose Data, let List be L3 or the list in which you entered the differences (don't forget to enter the differences!) and let Freq be 1.

   - If you choose Stats, enter the mean, SD, and sample size of the differences.

5. Let C-Level be the desired confidence level.

6. Choose Calculate and hit ENTER, which returns:

(__,__)	the confidence interval for the mean of the differences
x̄	the sample mean of the differences
Sx	the sample SD of the differences
n	the number of differences in the sample

CASIO FX-9750GII: MATCHED PAIRS T-INTERVAL

1. Compute the paired differences of the observations.

2. Using the computed differences, follow the instructions for a 1-sample t-interval.

GUIDED PRACTICE 7.27

(G)

In our UCLA textbook example, we had 68 paired differences. Because $df = 67$ was not on our t-table, we rounded the df down to 60. This gave us a 95% confidence interval (0.325, 6.834). Use a calculator to find the more exact 95% confidence interval based on 67 degrees of freedom. How different is it from the one we calculated based on 60 degrees of freedom?[10]

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

---

[9]No. This interval is attempting to capture the *average* increase. It is not attempting to capture 95% of the increases. Looking at Figure 7.15, we see that only a small percent had increases between 105.2 and 166.6.

[10]Choose TInterval or equivalent. We do not have all the data, so choose Stats on a TI or Var on a Casio. Enter x̄ = 3.58 and Sx = 13.42. Let n = 68 and C-Level = 0.95. This should give the interval (0.332, 6.828). The intervals are equivalent when rounded to two decimal places.

Section summary

- Paired data can come from a random sample or a matched pairs experiment. With paired data, we are often interested in whether the *difference* is positive, negative, or zero. For example, the difference of paired data from a matched pairs experiment tells us whether one treatment did better, worse, or the same as the other treatment for each subject.

- We use the notation $\bar{x}_{diff}$ to represent the mean of the sample differences. Likewise, s_{diff} is the standard deviation of the sample differences, and n_{diff} is the number of sample differences.

- To carry out inference on paired data, we first find all of the sample differences. Then, we perform a one-sample procedure on the *differences*. For this reason, the confidence interval and hypothesis test for paired data use the same t-procedures as the one-sample methods, where the degrees of freedom is given by $n_{diff} - 1$.

- When there is paired data and the parameter of interest is a mean of the differences:

  - Estimate μ_{diff} at the C% confidence level using a **matched pairs t-interval**.
  - Test H_0: $\mu_{diff} = 0$ at the α significance level using a **matched pairs t-test**.

- The conditions for the matched pairs t-interval and t-test are the same.

  1. There is paired data from a random sample or a matched pairs experiment.
  2. $n_{diff} \geq 30$ or the population of differences is nearly normal.

     If the number of differences is less than 30 and it is not known that the population of differences is nearly normal, we argue that the population of differences could be nearly normal if there is no strong skew or outliers in the sample differences.

- When the conditions are met, we calculate the confidence interval and the test statistic as we did in the previous section. Here, our data is a list of differences.

  Confidence interval: point estimate $\pm\ t^{\star} \times SE$ of estimate

  Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the mean of sample differences: $\bar{x}_{diff}$.

  The SE of estimate is the SE of a mean of sample differences: $\frac{s_{diff}}{\sqrt{n_{diff}}}$.

  The degrees of freedom is given by $df = n_{diff} - 1$.

Exercises

7.15   Air quality. Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years. Should we use a paired or non-paired test? Explain your reasoning.

7.16   True / False: paired. Determine if the following statements are true or false. If false, explain.

(a) In a paired analysis we first take the difference of each pair of observations, and then we do inference on these differences.

(b) Two data sets of different sizes cannot be analyzed as paired data.

(c) Consider two sets of data that are paired with each other. Each observation in one data set has a natural correspondence with exactly one observation from the other data set.

(d) Consider two sets of data that are paired with each other. Each observation in one data set is subtracted from the average of the other data set's observations.

7.17   Paired or not? Part I. In each of the following scenarios, determine if the data are paired.
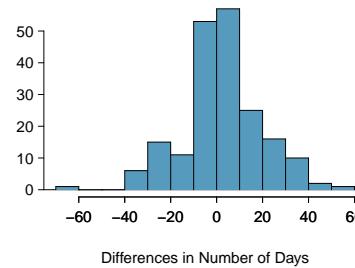
(a) Compare pre- (beginning of semester) and post-test (end of semester) scores of students.

(b) Assess gender-related salary gap by comparing salaries of randomly sampled men and women.

(c) Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E for the same group of patients.

(d) Assess effectiveness of a diet regimen by comparing the before and after weights of subjects.

7.18   Paired or not? Part II. In each of the following scenarios, determine if the data are paired.

(a) We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days.

(b) We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.

(c) A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.
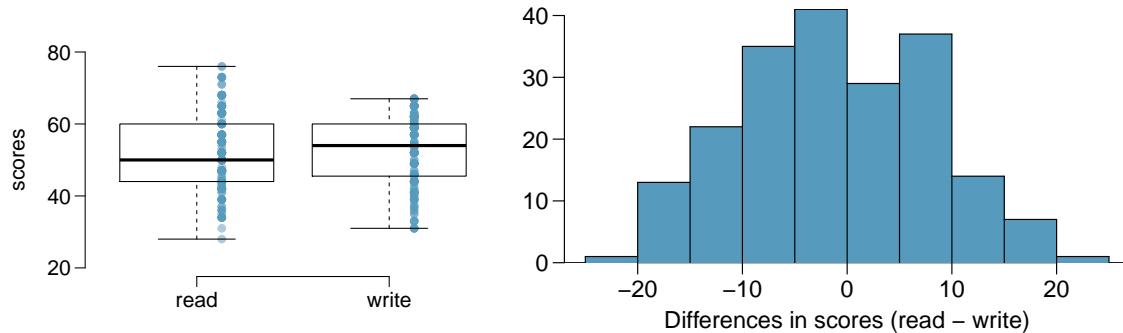
7.19   Global warming, Part I. Let's consider a limited set of climate data, examining temperature differences in 1948 vs 2018. We randomly sampled 197 locations from the National Oceanic and Atmospheric Administration's (NOAA) historical data, where the data was available for both years of interest. We want to know: were there more days with temperatures exceeding 90°F in 2018 or in 1948?[11] The difference in number of days exceeding 90°F (number of days in 2018 - number of days in 1948) was calculated for each of the 197 locations. The average of these differences was 2.9 days with a standard deviation of 17.2 days. We are interested in determining whether these data provide strong evidence that there were more days in 2018 that exceeded 90°F from NOAA's weather stations.

(a) Is there a relationship between the observations collected in 1948 and 2018? Or are the observations in the two groups independent? Explain.

(b) Write hypotheses for this research in symbols and in words.

(c) Check the conditions required to complete this test. A histogram of the differences is given to the right.

(d) Calculate the test statistic and find the p-value.

(e) Use $\alpha = 0.05$ to evaluate the test, and interpret your conclusion in context.

(f) What type of error might we have made? Explain in context what the error means.

Differences in Number of Days

7.20   High School and Beyond, Part I. The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

(a) Is there a clear difference in the average reading and writing scores?

(b) Are the reading and writing scores of each student independent of each other?

(c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

(d) Check the conditions required to complete this test.

(e) The average observed difference in scores is $\bar{x}_{read-write} = -0.545$, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?

(f) What type of error might we have made? Explain what the error means in the context of the application.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

7.21   Global warming, Part II. We considered the change in the number of days exceeding 90°F from 1948 and 2018 at 197 randomly sampled locations from the NOAA database in Exercise 7.19. The mean and standard deviation of the reported differences are 2.9 days and 17.2 days.

(a) Calculate a 90% confidence interval for the average difference between number of days exceeding 90°F between 1948 and 2018. We've already checked the conditions for you.

(b) Interpret the interval in context.

(c) Does the confidence interval provide convincing evidence that there were more days exceeding 90°F in 2018 than in 1948 at NOAA stations? Explain.

7.22   High school and beyond, Part II. We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey in Exercise 7.20. The mean and standard deviation of the differences are $\bar{x}_{read-write} = -0.545$ and 8.887 points.

(a) Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.

(b) Interpret this interval in context.

(c) Does the confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

7.3   Inference for the difference of two means

Often times we wish to compare two groups to each other to answer questions such as the following:

- Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack?

- Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

- Is there statistically significant evidence that one variation of an exam is harder than another variation?

- Are faculty willing to pay someone named "John" more than someone named "Jennifer"? If so, how much more?

Learning objectives

1. Determine when it is appropriate to use a paired t-procedure versus a two-sample t-procedure.

2. State and verify whether or not the conditions for inference on the difference of two means using the t-distribution are met.

3. Be able to use a calculator or other software to find the degrees of freedom associated with a two-sample t-procedure.

4. Carry out a complete confidence interval procedure for the difference of two means.

5. Carry out a complete hypothesis test for the difference of two means.

7.3.1 Sampling distribution for the difference of two means

Previously we explored the sampling distribution for the difference of two proportions. Here we consider the sampling distribution for the difference of two means. We are interested in the distribution of $\bar{x}_1 - \bar{x}_2$. We know that it is centered on $\mu_1 - \mu_2$. The standard deviation for the difference can be found as follows.

$$SD_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(SD_{\bar{x}_1}\right)^2 + \left(SD_{\bar{x}_2}\right)^2}$$
$$= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2}$$
$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Finally, we are interested in the shape of the sampling distribution of $\bar{x}_1 - \bar{x}_2$. It will be nearly normal when the sampling distribution of each of $\bar{x}_1$ and $\bar{x}_2$ are nearly normal.

7.3.2 Checking conditions for inference on a difference of means

When comparing two means, we carry out inference on a difference of means, $\mu_1 - \mu_2$. We will use the t-distribution just as we did when carrying out inference on a single mean. The assumptions are that the observations are independent, both between groups and within groups and that the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is nearly normal. We check whether these assumptions are reasonable by verifying the following conditions.

Independent. Observations can be considered independent when the data are collected from two independent random samples or, in the context of experiments, from two randomly assigned treatments. Randomly assigning subjects to treatments is equivalent to randomly assigning treatments to subjects.

Nearly normal sampling distribution. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be nearly normal when the sampling distribution of $\bar{x}_1$ and of $\bar{x}_2$ are nearly normal, that is when both population distributions are nearly normal or both sample sizes are at least 30.

As before, if the sample sizes are small and the population distributions are not known to be nearly normal, we look at the data for excessive skew or outliers. If we do no find excessive skew or outliers in either group, we consider the assumption that the populations are nearly normal to be reasonable.

7.3.3 Confidence intervals for a difference of means

What's in a name? Are employers more likely to offer interviews or higher pay to prospective employees when the name on a resume suggests the candidate is a man versus a woman? This is a challenging question to tackle, because employers are influenced by many aspects of a resume. Thinking back to Chapter 1 on data collection, we could imagine a host of confounding factors associated with name and gender. How could we possibly isolate just the factor of name? We would need an experiment in which name was the only variable and everything else was held constant.

Researchers at Yale carried out precisely this experiment. Their results were published in the Proceedings of the National Academy of Sciences (PNAS).[12] The researchers sent out resumes to faculty at academic institutions for a lab manager position. The resumes were identical, except that on half of them the applicant's name was John and on the other half, the applicant's name was Jennifer. They wanted to see if faculty, specifically faculty trained in conducting scientifically objective research, held implicit gender biases.

Unlike in the matched pairs scenario, each faculty member received only one resume. We are interested in comparing the mean salary offered to John relative to the mean salary offered to Jennifer. Instead of taking the average of a set of paired differences, we find the average of each group separately and take their difference. Let

$$\bar{x}_1 : \text{mean salary offered to John}$$
$$\bar{x}_2 : \text{mean salary offered to Jennifer}$$

We will use $\bar{x}_1 - \bar{x}_2$ as our point estimate for $\mu_1 - \mu_2$. The data is given in the table below.

Name	n	$\bar{x}$	s
John	63	$30,238	$5567
Jennifer	64	$26,508	$7247

We can calculate the difference as

$$\bar{x}_1 - \bar{x}_2 = 30,238 - 26,508 = 3730.$$

EXAMPLE 7.28

Interpret the point estimate 3730. Why might we want to construct a confidence interval?

The average salary offered to John was $3,730 higher than the average salary offered to Jennifer. Because there is randomness in which faculty ended up in the John group and which faculty ended up in the Jennifer group, we want to see if the difference of $3,730 is beyond what could be expected by random variation. In order to answer this, we will first want to calculate the SE for the difference of sample means.

EXAMPLE 7.29

Calculate and interpret the SE for the difference of sample means.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(5567)^2}{63} + \frac{(7247)^2}{64}} = 1151$$

The typical error in our estimate of $\mu_1 - \mu_2$, the real difference in mean salary that the faculty would offer John versus Jennifer, is $1151.

---

[12]https://www.pnas.org/content/109/41/16474

We see that the difference of sample means of \$3,730 is more than 3 SE above 0, which makes us think that the difference being 0 is unreasonable. We would like to construct a 95% confidence interval for the theoretical difference in mean salary that would be offered to John versus Jennifer. For this, we need the degrees of freedom associated with a two-sample t-interval.

For the one-sample t-procedure, the degrees of freedom is given by the simple expression $n-1$, where n is the sample size. For the two-sample t-procedures, however, there is a complex formula for calculating the degrees of freedom, which is based on the two sample sizes and the two sample standard deviations. In practice, we find the degrees of freedom using software or a calculator (see Section 7.3.4). If this is not possible, the alternative is to use the smaller of $n_1 - 1$ and $n_2 - 1$.

DEGREES OF FREEDOM FOR TWO-SAMPLE T-PROCEDURES

Use statistical software or a calculator to compute the degrees of freedom for two-sample t-procedures. If this is not possible, use the smaller of $n_1 - 1$ and $n_2 - 1$.

EXAMPLE 7.30

Verify that conditions are met for a two-sample t-test. Then, construct the 95% confidence interval for the difference of means.

We noted previously that this is an experiment and that the two treatments (name Jennifer and name John) were randomly assigned. Also, both sample sizes are well over 30, so the distribution of $\bar{x}_1 - \bar{x}_2$ is nearly normal. Using a calculator, we find that $df = 114.4$. Since 114 is not on the t-table, we round the degrees of freedom down to 100.[13] Using a t-table at row $df = 100$ with 95% confidence, we get a $t^\star = 1.984$. We calculate the confidence interval as follows.

$$\text{point estimate} \ \pm \ t^\star \times SE \text{ of estimate}$$
$$3730 \ \pm \ 1.984 \times 1151$$
$$= 3730 \ \pm \ 2284$$
$$= (1446, 6014)$$

Based on this interval, we are 95% confident that the true difference in mean salary that these faculty would offer John versus Jennifer is between \$1,495 and \$6,055. That is, we are 95% confident that the mean salary these faculty would offer John for a lab manager position is between \$1,446 and \$6,014 *more* than the mean salary they would offer Jennifer for the position.

The results of these studies and others like it are alarming and disturbing.[14] One aspect that makes this bias so difficult to address is that the experiment, as well-designed as it was, cannot send us much signal about *which* faculty are discriminating. Each faculty member received only one of the resumes. A faculty member that offered "Jennifer" a very low salary may have also offered "John" a very low salary.

We might imagine an experiment in which each faculty received both resumes, so that we could compare how much they would offer a Jennifer versus a John. However, the matched pairs scenario is clearly not possible in this case, because what makes the experiment work is that the resumes are *exactly the same* except for the name. An employer would notice something fishy if they received two identical resumes. It is only possible to say that overall, the faculty were willing to offer John more money for the lab manager position than Jennifer. Finding proof of bias for individual cases is a persistent challenge in enforcing anti-discrimination laws.

---

[13]Using technology, we get a more precise interval, based on 114.4 df: (1495, 6055).

[14]A similar study sent out identical resumes with different names to investigate the importance of perceived race. Resumes with a name commonly perceived to be for a White person (e.g. Emily) were 50% more likely to receive a callback than the same resume with a name commonly perceived to be for a Black person (e.g. Lakisha). https://www.nber.org/papers/w9873

CONSTRUCTING A CONFIDENCE INTERVAL FOR THE DIFFERENCE OF TWO MEANS

To carry out a complete confidence interval procedure to estimate the difference of two means $\mu_1 - \mu_2$,

Identify: Identify the parameter and the confidence level, C%.

The parameter will be a difference of means, e.g. the true difference in mean cholesterol reduction (mean treatment A − mean treatment B).

Choose: Choose the appropriate interval procedure and identify it by name.

Here we choose the **2-sample t-interval**.

Check: Check conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to be nearly normal.

1. Data come from 2 independent random samples or 2 randomly assigned treatments.
2. $n_1 \geq 30$ and $n_2 \geq 30$ or both population distributions are nearly normal.

   If the sample sizes are less than 30 and the population distributions are unknown, check for strong skew or outliers in either data set. If neither is found, the condition that both population distributions are nearly normal is considered reasonable.

Calculate: Calculate the confidence interval and record it in interval form.

point estimate $\pm\ t^\star \times SE$ of estimate, df: use calculator or other technology

point estimate: the difference of sample means $\bar{x}_1 - \bar{x}_2$

SE of estimate: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$t^\star$: use a t-table at row df and confidence level C%

(___, ___)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

We are C% confident that the true *difference in mean* [...] is between ____ and ____. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.

EXAMPLE 7.31

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, she shuffled the exams together to ensure each student received a random version. Summary statistics for how students performed on these two exams are shown in Figure 7.31. Anticipating complaints from students who took Version B, she would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A. Use a 95% confidence interval to estimate the difference in average score: version A - version B.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	30	74.1	20	32	100

Identify: The parameter we want to estimate is $\mu_1 - \mu_2$, which is the true average score under Version A $-$ the true average score under Version B. We will estimate this parameter at the 95% confidence level.

Choose: Because we are comparing two means, we will use a 2-sample t-interval.

Check: The data was collected from an experiment with two randomly assigned treatments, Version A and Version B of test. Both groups sizes are 30, so the condition that they are at least 30 is met.

Calculate: We will calculate the confidence interval as follows.

$$\text{point estimate} \ \pm \ t^\star \times SE \text{ of estimate}$$

The point estimate is the difference of sample means: $\bar{x}_1 - \bar{x}_2 = 79.4 - 74.1 = 5.3$

The SE of a difference of sample means is: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{30}} = 4.46$

In order to find the critical value $t^\star$, we must first find the degrees of freedom. Using a calculator, we find $df = 51.9$. We round down to 50, and using a t-table at row $df = 50$ and confidence level 95%, we get $t^\star = 2.009$.

The 95% confidence interval is given by:

$$(79.4 - 74.1) \pm 2.009 \times \sqrt{\frac{14^2}{30} + \frac{20^2}{30}} \qquad df = 45.97$$
$$5.3 \pm 2.009 \times 4.46$$
$$= (-3.66, \ 14.26)$$

Conclude: We are 95% confident that the true difference in average score between Version A and Version B is between -2.5 and 13.1 points. Because the interval contains both positive and negative values, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

7.3.4   Calculator: the 2-sample t-interval

▶◄ TI-83/84: 2-SAMPLE T-INTERVAL

Use `STAT`, `TESTS`, `2-SampTInt`.

1. Choose `STAT`.

2. Right arrow to `TESTS`.

3. Down arrow and choose `0:2-SampTTInt`.

4. Choose `Data` if you have all the data or `Stats` if you have the means and standard deviations.

    - If you choose Data, let `List1` be `L1` or the list that contains sample 1 and let `List2` be `L2` or the list that contains sample 2 (don't forget to enter the data!). Let `Freq1` and `Freq2` be `1`.

    - If you choose `Stats`, enter the mean, SD, and sample size for sample 1 and for sample 2.

5. Let `C-Level` be the desired confidence level and let `Pooled` be `No`.

6. Choose `Calculate` and hit `ENTER`, which returns:

`(__,__)`	the confidence interval	`Sx1`	SD of sample 1
`df`	degrees of freedom	`Sx2`	SD of sample 2
$\bar{x}_1$	mean of sample 1	`n1`	size of sample 1
$\bar{x}_2$	mean of sample 2	`n2`	size of sample 2

▶◄ CASIO FX-9750GII: 2-SAMPLE T-INTERVAL

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).

2. If necessary, enter the data into a list.

3. Choose the `INTR` option (`F4` button).

4. Choose the `t` option (`F2` button).

5. Choose the `2-S` option (`F2` button).

6. Choose either the `Var` option (`F2`) or enter the data in using the `List` option.

7. Specify the test details:

    - Confidence level of interest for `C-Level`.

    - If using the `Var` option, enter the summary statistics for each group. If using `List`, specify the lists and leave `Freq` values at `1`.

    - Choose whether to pool the data or not.

8. Hit the `EXE` button, which returns

`Left`, `Right`	ends of the confidence interval
`df`	degrees of freedom
$\bar{x}1$, $\bar{x}2$	sample means
`sx1`, `sx2`	sample standard deviations
`n1`, `n2`	sample sizes

GUIDED PRACTICE 7.32

Use the data below and a calculator to find a 95% confidence interval for the difference in average scores between Version A and Version B of the exam from the previous example.[15]

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	30	74.1	20	32	100

7.3.5 Hypothesis testing for the difference of two means

Four cases from a data set called `ncbirths`, which represents mothers and their newborns in North Carolina, are shown in Figure 7.16. We are particularly interested in two variables: `weight` and `smoke`. The `weight` variable represents the weights of the newborns and the `smoke` variable describes which mothers smoked during pregnancy. We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? The smoking group includes a random sample of 50 cases and the nonsmoking group contains a random sample of 100 cases, represented in Figure 7.17.

	fAge	mAge	weeks	weight	sex	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Figure 7.16: Four cases from the `ncbirths` data set. The value "NA", shown for the first two entries of the first variable, indicates pieces of data that are missing.

Newborn Weights (lbs) From Mothers Who Smoked

Newborn Weights (lbs) From Mothers Who Did Not Smoke
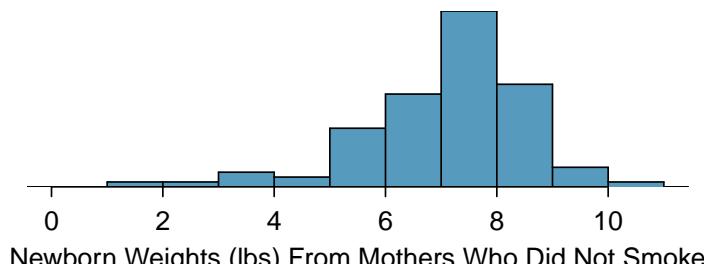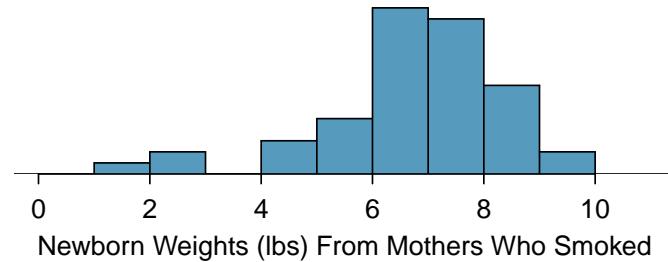
Figure 7.17: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

---

[15]Choose `2-SampTInt` or equivalent. Because we have the summary statistics rather than all of the data, choose `Stats`. Let `x̄1`=79.41, `Sx1`=14, `n1`=30, `x̄2`=74.1, `Sx2` = 20, and `n2` = 30. The interval is $(-3.6, 14.2)$ with $df = 51.9$.

EXAMPLE 7.33

Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

———

Let μ_1 represent the mean for mothers that did smoke and μ_2 represent the mean for mothers that did not smoke. We will take the difference as: smoker − nonsmoker. The null hypothesis represents the case of no difference between the groups.

H_0: $\mu_1 - \mu_2 = 0$. There is no difference in average birth weight for newborns from mothers who did and did not smoke.

H_A: $\mu_1 - \mu_2 \neq 0$. There is some difference in average newborn weights from mothers who did and did not smoke.

We check the two conditions necessary to use the t-distribution to the difference in sample means. (1) Because the data come from a sample, we need there to be two independent random samples. In fact, there was only one random sample, but it is reasonable that the two groups here are independent of each other, so we will consider the assumption of independence reasonable. (2) The sample sizes of 50 and 100 are well over 30, so we do not worry about the distributions of the original populations. Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Figure 7.18: Summary statistics for the `ncbirths` data set.

EXAMPLE 7.34

We will use the summary statistics in Figure 7.18 for this exercise.
(a) What is the point estimate of the population difference, $\mu_1 - \mu_2$?
(b) Compute the standard error of the point estimate from part (a).

———

(a) The point estimate is the difference of sample means: $\bar{x}_1 - \bar{x}_2 = 6.78 - 7.18 = -0.40$ pounds.
(b) The standard error for a difference of sample means is calculated analogously to the standard deviation for a difference of sample means.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}} = \sqrt{\frac{1.43^2}{100} + \frac{1.60^2}{50}} = 0.26 \text{ pounds}$$

EXAMPLE 7.35

Compute the test statistic.

———
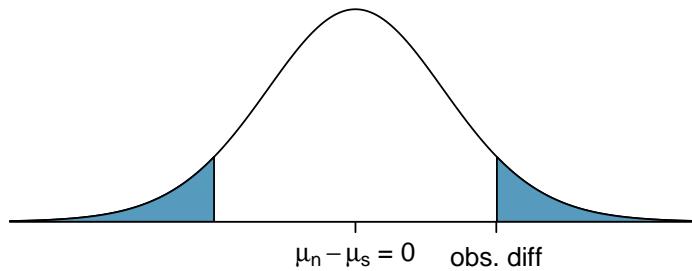
We have already found the point estimate and the SE of estimate. The null hypothesis is that the two means are equal, or that their difference equals 0. The null value for the difference, therefore is 0. We now have everything we need to compute the test statistic.

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}} = \frac{0.40 - 0}{0.26} = 1.54$$

EXAMPLE 7.36

Draw a picture to represent the p-value for this hypothesis test, then calculate the p-value.

To depict the p-value, we draw the distribution of the point estimate as though H_0 were true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.

$$\mu_n - \mu_s = 0 \quad \text{obs. diff}$$

We saw previously that the degrees of freedom can be found using software or using the smaller of $n_1 - 1$ and $n_2 - 1$. If we use $50 - 1 = 49$ degrees of freedom, we find that the area in the upper tail is 0.065. The p-value is twice this, or $2 \times 0.065 = 0.130$. See Section 7.3.6 for a shortcut to compute the degrees of freedom and p-value on a calculator.

EXAMPLE 7.37

What can we conclude from this p-value? Use a significance level of $\alpha = 0.05$.

This p-value of 0.130 is larger the significance level of 0.05, so we do not reject the null hypothesis. There is not sufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

EXAMPLE 7.38

Does the conclusion to Example 7.35 mean that smoking and average birth weight are unrelated?

Not necessarily. It is possible that there is some difference but that we did not detect it. The result must be considered in light of other evidence and research. In fact, larger data sets do tend to show that women who smoke during pregnancy have smaller newborns.

GUIDED PRACTICE 7.39

If we made an error in our conclusion, which type of error could we have made: Type I or Type II?[16]

GUIDED PRACTICE 7.40

If we made a Type II Error and there is a difference, what could we have done differently in data collection to be more likely to detect the difference?[17]

---

[16]Since we did not reject H_0, it is possible that we made a Type II Error. It is possible that there is some difference but that we did not detect it.

[17]We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists. In other words, increasing the sample size increases the power of the test.

HYPOTHESIS TEST FOR THE DIFFERENCE OF TWO MEANS

To carry out a complete hypothesis test to test the claim that two means μ_1 and μ_2 are equal to each other,

Identify: Identify the hypotheses and the significance level, α.

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$; H_A: $\mu_1 > \mu_2$;   or H_A: $\mu_1 < \mu_2$

Choose: Choose the appropriate test procedure and identify it by name.

Here we choose the **2-sample t-test**.

Check: Check conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to be nearly normal.

1. Data come from 2 independent random samples or 2 randomly assigned treatments.
2. $n_1 \geq 30$ and $n_2 \geq 30$ or both population distributions are nearly normal.

   If the sample sizes are less than 30 and the population distributions are unknown, check for excessive skew or outliers in either data set. If neither is found, the condition that both population distributions are nearly normal is considered reasonable.

Calculate: Calculate the t-statistic, df, and p-value.

$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$ df: use calculator or other technology

point estimate: the difference of sample means $\bar{x}_1 - \bar{x}_2$

SE of estimate: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

p-value = (based on the t-statistic, the df, and the direction of H_A)
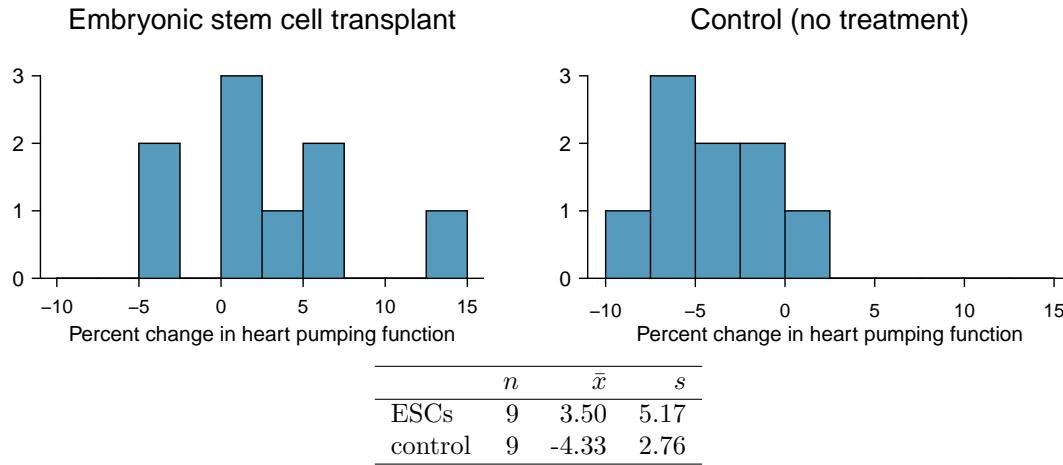
Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].

EXAMPLE 7.41

Do embryonic stem cells (ESCs) help improve heart function following a heart attack? The following table and figure summarize results from an experiment to test ESCs in sheep that had a heart attack.

Embryonic stem cell transplant

Control (no treatment)

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured. A positive value generally corresponds to increased pumping capacity, which suggests a stronger recovery. The sample data is also graphed. Use the given information and an appropriate statistical test to answer the research question.

Identify: Let μ_1 be the mean percent change for sheep that receive ESC and let μ_2 be the mean percent change for sheep in the control group. We will use an $\alpha = 0.05$ significance level.
H_0: $\mu_1 - \mu_2 = 0$. The stem cells do not improve heart pumping function.
H_A: $\mu_1 - \mu_2 > 0$. The stem cells do improve heart pumping function.

Choose: Because we are hypothesizing about a difference of means we choose the 2-sample t-test.

Check: The data come from an experiment with two randomly assigned treatments. The group sizes are small, but the data show no excessive skew or outliers, so the assumption that the population distributions are normal is reasonable.

Calculate: We will calculate the t-statistic and the p-value.

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

The point estimate is the difference of sample means: $\bar{x}_1 - \bar{x}_2 = 3.50 - (-4.33) = 7.83$

The SE of a difference of sample means: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(5.17)^2}{9} + \frac{(2.76)^2}{9}} = 1.95$

$$T = \frac{3.50 - (-4.33) - 0}{\sqrt{\frac{(5.17)^2}{9} + \frac{(2.76)^2}{9}}} = \frac{7.83 - 0}{1.95} = 4.01$$

Because H_A is an upper tail test ($>$), the p-value corresponds to the area to the right of $t = 4.01$ with the appropriate degrees of freedom. Using a calculator, we find get $df = 12.2$ and p-value $= 8.4 \times 10^{-4} = 0.00084$.

Conclude: The p-value is much less than 0.05, so we reject the null hypothesis. There is sufficient evidence that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack.

7.3.6   Calculator: the 2-sample t-test

TI-83/84: 2-SAMPLE T-TEST

Use `STAT`, `TESTS`, `2-SampTTest`.

1. Choose `STAT`.
2. Right arrow to `TESTS`.
3. Choose `4:2-SampTTest`.
4. Choose `Data` if you have all the data or `Stats` if you have the means and standard deviations.
   - If you choose `Data`, let `List1` be `L1` or the list that contains sample 1 and let `List2` be `L2` or the list that contains sample 2 (don't forget to enter the data!). Let `Freq1` and `Freq2` be `1`.
   - If you choose `Stats`, enter the mean, SD, and sample size for sample 1 and for sample 2
5. Choose $\neq$, $<$, or $>$ to correspond to H_A.
6. Let `Pooled` be `NO`.
7. Choose `Calculate` and hit `ENTER`, which returns:

t	t statistic	Sx1	SD of sample 1
p	p-value	Sx2	SD of sample 2
df	degrees of freedom	n1	size of sample 1
$\bar{x}_1$	mean of sample 1	n2	size of sample 2
$\bar{x}_2$	mean of sample 2		

CASIO FX-9750GII: 2-SAMPLE T-TEST

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).
2. If necessary, enter the data into a list.
3. Choose the `TEST` option (`F3` button).
4. Choose the `t` option (`F2` button).
5. Choose the `2-S` option (`F2` button).
6. Choose either the `Var` option (`F2`) or enter the data in using the `List` option.
7. Specify the test details:
   - Specify the sidedness of the test using the `F1`, `F2`, and `F3` keys.
   - If using the `Var` option, enter the summary statistics for each group. If using `List`, specify the lists and leave `Freq` values at `1`.
   - Choose whether to pool the data or not.
8. Hit the `EXE` button, which returns

$\mu1 \_\_ \mu2$	alt. hypothesis	$\bar{x}1$, $\bar{x}2$	sample means
t	t statistic	sx1, sx2	sample standard deviations
p	p-value	n1, n2	sample sizes
df	degrees of freedom		

GUIDED PRACTICE 7.42

Use the data below and a calculator to find the test statistics and p-value for a one-sided test, testing whether there is evidence that embryonic stem cells (ESCs) help improve heart function for sheep that have experienced a heart attack.[18]

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

---

[18]Choose `2-SampTTest` or equivalent. Because we have the summary statistics rather than all of the data, choose `Stats`. Let `x̄1`=3.50, `Sx1`=5.17, `n1`=9, `x̄2`=-4.33, `Sx2` = 2.76, and `n2` = 9. We get `t` = 4.01, and the p-value $p = 8.4 \times 10^{-4} = 0.00084$. The degrees of freedom for the test is `df` = 12.2.

Section summary

- This section introduced inference for a difference of means, which is distinct from inference for a mean of differences. To calculate a difference of means, $\bar{x}_1 - \bar{x}_2$, we first calculate the mean of each group, then we take the difference between those two numbers. To calculate a mean of difference, $\bar{x}_{diff}$, we first calculate all of the differences, then we find the mean of those differences.

- Inference for a difference of means is based on the t-distribution. The degrees of freedom is complicated to calculate and we rely on a calculator or other software to calculate this.[19]

- When there are two samples or treatments and the parameter of interest is a difference of means:
    - Estimate $\mu_1 - \mu_2$ at the C% confidence level using a **2-sample t-interval**.
    - Test H_0: $\mu_1 - \mu_2 = 0$ (i.e. $\mu_1 = \mu_2$) at the α significance level using a **2-sample t-test** .

- The conditions for the two sample t-interval and t-test are the same.

    1. The data come from 2 independent random samples or 2 randomly assigned treatments.

    2. $n_1 \geq 30$ and $n_2 \geq 30$ or both population distributions are nearly normal.

        If the sample sizes are less than 30 and it is not known that both population distributions are nearly normal, check for excessive skew or outliers in the data. If neither exists, the condition that both population distributions could be nearly normal is considered reasonable.

- When the conditions are met, we calculate the confidence interval and the test statistic as follows.

    Confidence interval: point estimate $\pm\ t^{\star} \times SE$ of estimate

    Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

    Here the point estimate is the difference of sample means: $\bar{x}_1 - \bar{x}_2$.
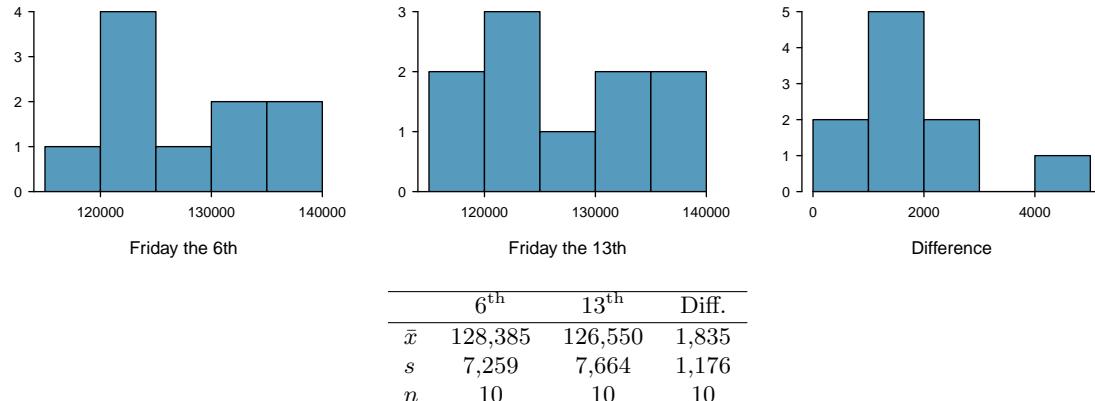
    The SE of estimate is the SE of a difference of sample means: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

   Find and record the df using a calculator or other software.

---

[19]If this is not available, one can use $df = min(n_1 - 1, n_2 - 1)$.

Exercises

7.23   Friday the 13$^{\text{th}}$, Part I. In the early 1990's, researchers in the UK collected data on traffic flow, number of shoppers, and traffic accident related emergency room admissions on Friday the 13$^{\text{th}}$ and the previous Friday, Friday the 6$^{\text{th}}$. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6$^{\text{th}}$ and Friday the 13$^{\text{th}}$ for many such date pairs. Also given are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th.[20]


	6$^{\text{th}}$	13$^{\text{th}}$	Diff.
$\bar{x}$	128,385	126,550	1,835
s	7,259	7,664	1,176
n	10	10	10
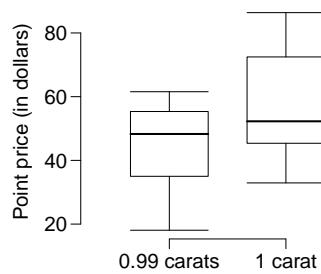
(a) Are there any underlying structures in these data that should be considered in an analysis? Explain.

(b) What are the hypotheses for evaluating whether the number of people out on Friday the 6$^{\text{th}}$ is different than the number out on Friday the 13$^{\text{th}}$?

(c) Check conditions to carry out the hypothesis test from part (b).

(d) Calculate the test statistic and the p-value.

(e) What is the conclusion of the hypothesis test?

(f) Interpret the p-value in this context.

(g) What type of error might have been made in the conclusion of your test? Explain.

7.24   Diamonds, Part I. Prices of diamonds are determined by what is known as the 4 Cs: cut, clarity, color, and carat weight. The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and compare the average prices of the diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. For a 1 carat diamond, we divide the price by 100. The distributions and some sample statistics are shown below.[21]

Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.
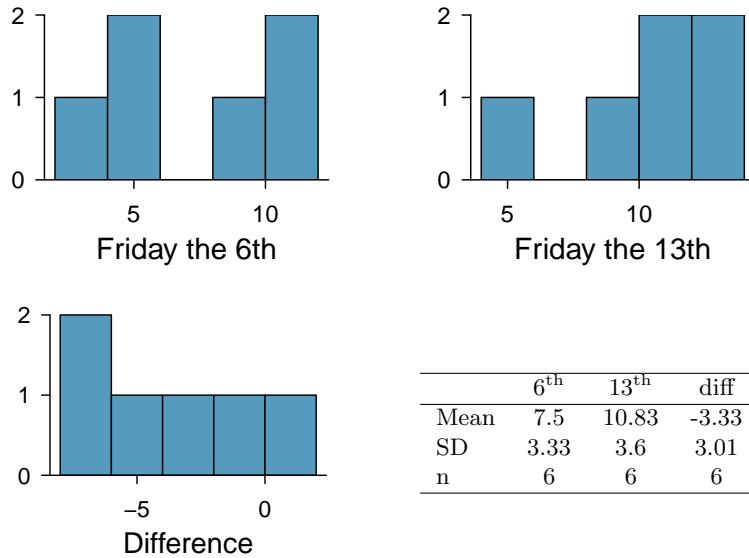
	0.99 carats	1 carat
Mean	$44.51	$56.81
SD	$13.32	$16.13
n	23	23

---

[20]T.J. Scanlon et al. "Is Friday the 13th Bad For Your Health?" In: *BMJ* 307 (1993), pp. 1584–1586.
[21]H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

7.25 Friday the 13th, Part II. The Friday the 13th study reported in Exercise 7.23 also provides data on traffic accident related emergency room admissions. The distributions of these counts from Friday the 6th and Friday the 13th are shown below for six such paired dates along with summary statistics. You may assume that conditions for inference are met.
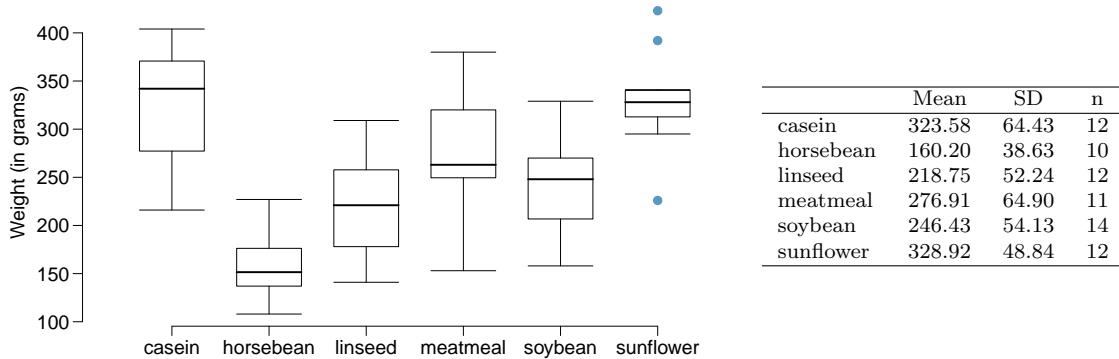
	6th	13th	diff
Mean	7.5	10.83	-3.33
SD	3.33	3.6	3.01
n	6	6	6

(a) Conduct a hypothesis test to evaluate if there is a difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

(b) Calculate a 95% confidence interval for the difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

(c) The conclusion of the original study states, "Friday 13th is unlucky for some. The risk of hospital admission as a result of a transport accident may be increased by as much as 52%. Staying at home is recommended." Do you agree with this statement? Explain your reasoning.

7.26 Diamonds, Part II. In Exercise 7.24, we discussed diamond prices (standardized by weight) for diamonds with weights 0. 99 carats and 1 carat. See the table for summary statistics, and then construct a 95% confidence interval for the average difference between the standardized prices of 0.99 and 1 carat diamonds. You may assume the conditions for inference are met.

	0.99 carats	1 carat
Mean	$44.51	$56.81
SD	$13.32	$16.13
n	23	23
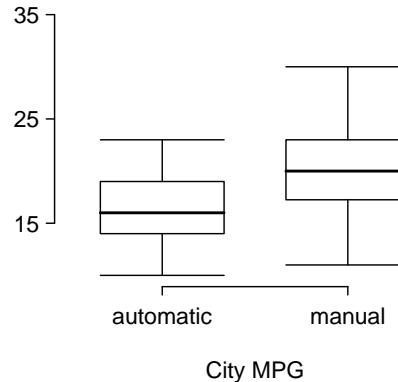
7.27   Chicken diet and weight, Part I. Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Below are some summary statistics from this data set along with box plots showing the distribution of weights by feed type.[22]

	Mean	SD	n
casein	323.58	64.43	12
horsebean	160.20	38.63	10
linseed	218.75	52.24	12
meatmeal	276.91	64.90	11
soybean	246.43	54.13	14
sunflower	328.92	48.84	12

(a) Describe the distributions of weights of chickens that were fed linseed and horsebean.

(b) Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different? Use a 5% significance level.

(c) What type of error might we have committed? Explain.

(d) Would your conclusion change if we used $\alpha = 0.01$?

7.28   Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[23]

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26

City MPG
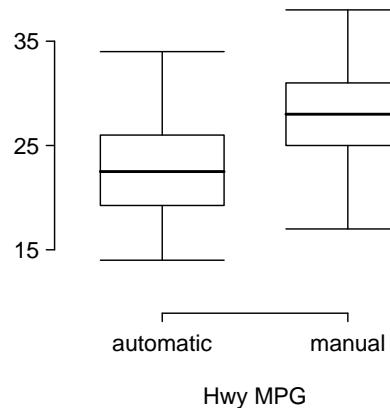
7.29   Chicken diet and weight, Part II. Casein is a common weight gain supplement for humans. Does it have an effect on chickens? Using data provided in Exercise 7.27, test the hypothesis that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean. If your hypothesis test yields a statistically significant result, discuss whether or not the higher average weight of chickens can be attributed to the casein diet. Assume that conditions for inference are satisfied.

---

[22]Chicken Weights by Feed Type, from the `datasets` package in R..
[23]U.S. Department of Energy, Fuel Economy Data, 2012 Datafile.

7.30   Fuel efficiency of manual and automatic cars, Part II. The table provides summary statistics on highway fuel economy of the same 52 cars from Exercise 7.28. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.[24]
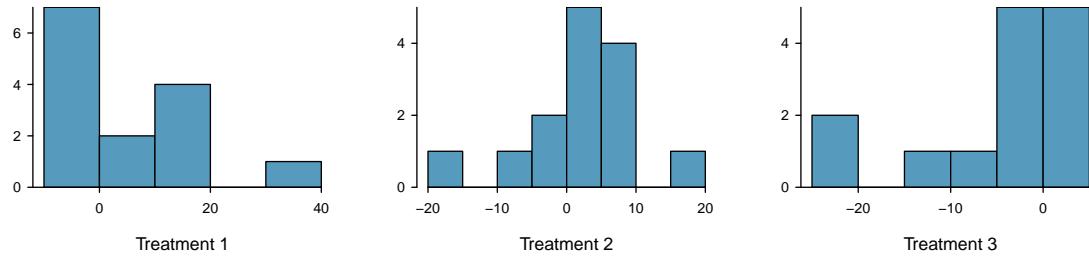
	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26

7.31   Prison isolation experiment, Part I. Subjects from Central Prison in Raleigh, NC, volunteered for an experiment involving an "isolation" experience. The goal of the experiment was to find a treatment that reduces subjects' psychopathic deviant T scores. This score measures a person's need for control or their rebellion against control, and it is part of a commonly used mental health test called the Minnesota Multiphasic Personality Inventory (MMPI) test. The experiment had three treatment groups:

(1) Four hours of sensory restriction plus a 15 minute "therapeutic" tape advising that professional help is available.

(2) Four hours of sensory restriction plus a 15 minute "emotionally neutral" tape on training hunting dogs.

(3) Four hours of sensory restriction but no taped message.

Forty-two subjects were randomly assigned to these treatment groups, and an MMPI test was administered before and after the treatment. Distributions of the differences between pre and post treatment scores (pre - post) are shown below, along with some sample statistics. Use this information to independently test the effectiveness of each treatment. Make sure to clearly state your hypotheses, check conditions, and interpret results in the context of the data.[25]

	Tr 1	Tr 2	Tr 3
Mean	6.21	2.86	-3.21
SD	12.3	7.94	8.57
n	14	14	14

7.32   True / False: comparing means. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

(a) When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal model for the difference in means since $n_2 \geq 30$.

(b) As the degrees of freedom increases, the t-distribution approaches normality.

(c) We use a pooled standard error for calculating the standard error of the difference between means when sample sizes of groups are equal to each other.

---

[24]U.S. Department of Energy, Fuel Economy Data, 2012 Datafile.
[25]Prison isolation experiment, stat.duke.edu/resources/datasets/prison-isolation.

Chapter highlights

We've reviewed a wide set of inference procedures over the last 3 chapters. Let's revisit each and discuss the similarities and differences among them. The following confidence intervals and tests are structurally the same – they all involve inference on a population parameter, where that parameter is a proportion, a difference of proportions, a mean, a mean of differences, or a difference of means.

- 1-proportion z-test/interval
- 2-proportion z-test/interval
- 1-sample t-test/interval
- matched pairs t-test/interval
- 2-sample t-test/interval

The above inferential procedures all involve a **point estimate**, a **standard error** of the estimate, and an assumption about the **shape of the sampling distribution** of the point estimate.

From Chapter 6, the χ^2 tests and their uses are as follows:

- χ^2 goodness of fit - compares a categorical variable to a known/fixed distribution.
- χ^2 test of homogeneity - compares a categorical variable across multiple groups.
- χ^2 test of independence - looks for association between two categorical variables.

χ^2 is a measure of *overall* deviation between observed values and expected values. These tests stand apart from the others because when using χ^2 there is not a parameter of interest. For this reason there are no confidence intervals using χ^2. Also, for χ^2 tests, the hypotheses are usually written in words, because they are about the *distribution* of one or more categorical variables, not about a single parameter.

While formulas and conditions vary, all of these procedures follow the same basic logic and process.

- For a confidence interval, identify the parameter to be estimated and the confidence level. For a hypothesis test, identify the hypotheses to be tested and the significance level.
- Choose the correct procedure.
- Check that both conditions for its use are met.
- Calculate the confidence interval or the test statistic and p-value, as well as the *df* if applicable.
- Interpret the results and draw a conclusion based on the data.

For a summary of these hypothesis test and confidence interval procedures (including one more that we will encounter in Section 8.3), see the Inference Guide in Appendix D.2.
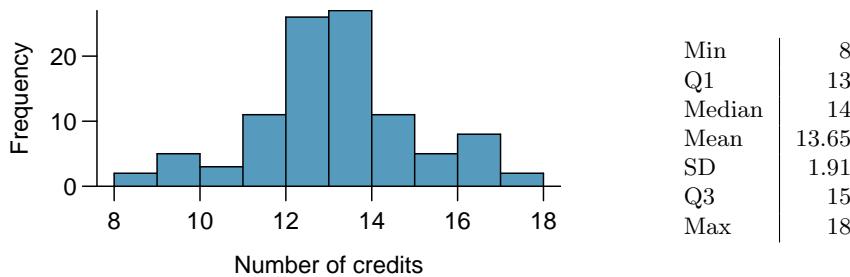
Chapter exercises

7.33   Gaming and distracted eating, Part I. A group of researchers are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption. To test this hypothesis, they monitored food intake for a group of 44 patients who were randomized into two equal groups. The treatment group ate lunch while playing solitaire, and the control group ate lunch without any added distractions. Patients in the treatment group ate 52.1 grams of biscuits, with a standard deviation of 45.1 grams, and patients in the control group ate 27.1 grams of biscuits, with a standard deviation of 26.4 grams. Do these data provide convincing evidence that the average food intake (measured in amount of biscuits consumed) is different for the patients in the treatment group? Assume that conditions for inference are satisfied.[26]

7.34   Gaming and distracted eating, Part II. The researchers from Exercise 7.33 also investigated the effects of being distracted by a game on how much people eat. The 22 patients in the treatment group who ate their lunch while playing solitaire were asked to do a serial-order recall of the food lunch items they ate. The average number of items recalled by the patients in this group was 4. 9, with a standard deviation of 1.8. The average number of items recalled by the patients in the control group (no distraction) was 6.1, with a standard deviation of 1.8. Do these data provide strong evidence that the average number of food items recalled by the patients in the treatment and control groups are different?

7.35   Sample size and pairing. Determine if the following statement is true or false, and if false, explain your reasoning: If comparing means of two groups with equal sample sizes, always use a paired test.

7.36   College credits. A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.

Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

(a) What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

(c) Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning.

(d) The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

(e) The sample means given above are point estimates for the mean number of credits taken by all students at that college. What measures do we use to quantify the variability of this estimate? Compute this quantity using the data from the original sample.

[26]R.E. Oldham-Cooper et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake". In: *The American Journal of Clinical Nutrition* 93.2 (2011), p. 308.

7.37 Hen eggs. The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.
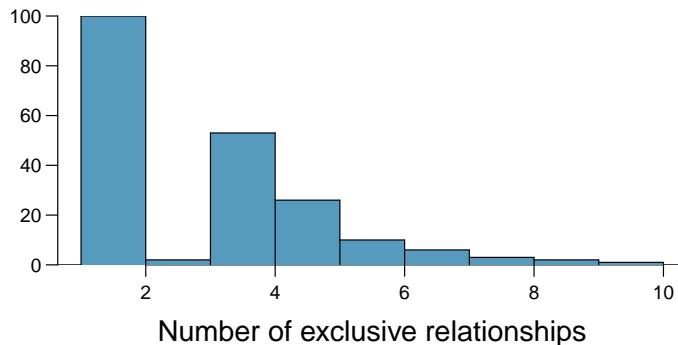
(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value.

(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

7.38 Forest management. Forest rangers wanted to better understand the rate of growth for younger trees in the park. They took measurements of a random sample of 50 young trees in 2009 and again measured those same trees in 2019. The data below summarize their measurements, where the heights are in feet:

	2009	2019	Differences
$\bar{x}$	12.0	24.5	12.5
s	3.5	9.5	7.2
n	50	50	50
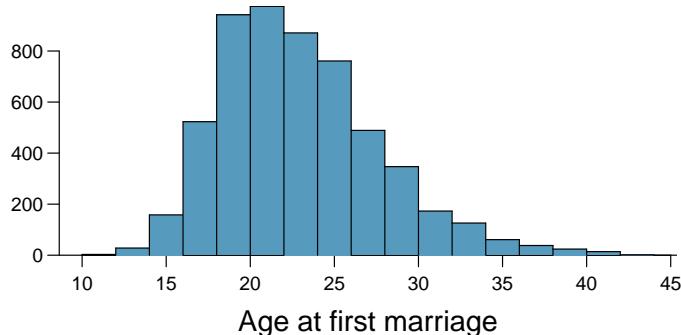
Construct a 99% confidence interval for the average growth of (what had been) younger trees in the park over 2009-2019.

7.39 Exclusive relationships. A survey conducted on a reasonably random sample of 203 undergraduates asked, among many other questions, about the number of exclusive relationships these students have been in. The histogram below shows the distribution of the data from this sample. The sample average is 3.2 with a standard deviation of 1.97.

Number of exclusive relationships

Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.

7.40   Age at first marriage, Part I. The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram below shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72.[27]

Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.

7.41   Online communication. A study suggests that the average college student spends 10 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and find that on average they spent 13.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} < 10 \ hours$$
$$H_A : \bar{x} > 13.5 \ hours$$

7.42   Age at first marriage, Part II. Exercise 7.40 presents the results of a 2006 - 2010 survey showing that the average age of women at first marriage is 23.44. Suppose a social scientist thinks this value has changed since the survey was taken. Below is how she set up her hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} \neq 23.44 \ years \ old$$
$$H_A : \bar{x} = 23.44 \ years \ old$$

---

[27]Centers for Disease Control and Prevention, National Survey of Family Growth, 2010.

Chapter 8

Introduction to linear regression

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with straight lines are overlaid on scatterplots. Linear models can be used to see trends and to make predictions.

For videos, slides, and other resources, please visit
www.openintro.org/ahss

8.1 Line fitting, residuals, and correlation

In this section, we investigate bivariate data. We examine criteria for identifying a linear model and introduce a new bivariate summary called *correlation*. We answer questions such as the following:

- How do we quantify the strength of the linear association between two numerical variables?

- What does it mean for two variables to have no association or to have a nonlinear association?

- Once we fit a model, how do we measure the error in the model's predictions?

Learning objectives

1. Distinguish between the data point y and the predicted value $\hat{y}$ based on a model.
2. Calculate a residual and draw a residual plot.
3. Interpret the standard deviation of the residuals.
4. Interpret the correlation coefficient and estimate it from a scatterplot.
5. Know and apply the properties of the correlation coefficient.

8.1.1 Fitting a line to data

Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker TGT, April 26th, 2012). We let x be the number of stocks to purchase and y be the total cost. Because the cost is computed using a linear formula, the linear fit is perfect, and the equation for the line is: $y = 5 + 57.49x$. If we know the number of stocks purchased, we can determine the cost based on this linear equation with no error. Additionally, we can say that each additional share of the stock cost $57.49 and that there was a $5 fee for the transaction.
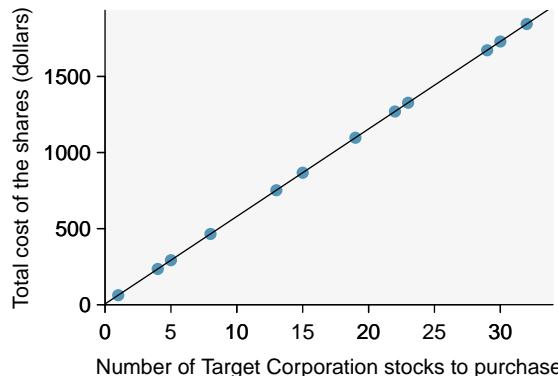

Figure 8.1: Total cost of a trade against number of shares purchased.

Perfect linear relationships are unrealistic in almost any natural process. For example, if we took family income (x), this value would provide some useful information about how much financial support a college may offer a prospective student (y). However, the prediction would be far from perfect, since other factors play a role in financial support beyond a family's income.

It is rare for all of the data to fall perfectly on a straight line. Instead, it's more common for data to appear as a *cloud of points*, such as those shown in Figure 8.2. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it.

In each of these examples, we can consider how to draw a "best fit line". For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less? As we move forward in this chapter, we will learn different criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.
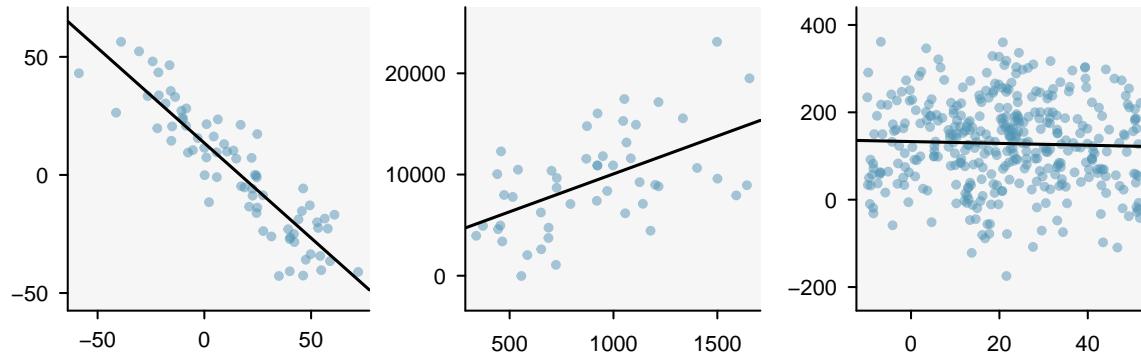
Figure 8.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

We will also see examples in this chapter where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 8.3 where there is a very strong relationship between the variables even though the trend is not linear.

Figure 8.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

8.1.2 Using linear regression to predict possum head lengths

Brushtail possums are a marsupial that lives in Australia. A photo of one is shown in Figure 8.4. Researchers captured 104 of these animals and took body measurements before releasing the animals back into the wild. We consider two of these measurements: the total length of each possum, from head to tail, and the length of each possum's head.

Figure 8.5 shows a scatterplot for the head length and total length of the 104 possums. Each point represents a single point from the data.

Figure 8.4: The common brushtail possum of Australia.

Photo by Peter Firminger on Flickr: http://flic.kr/p/6aPTn CC BY 2.0 license.
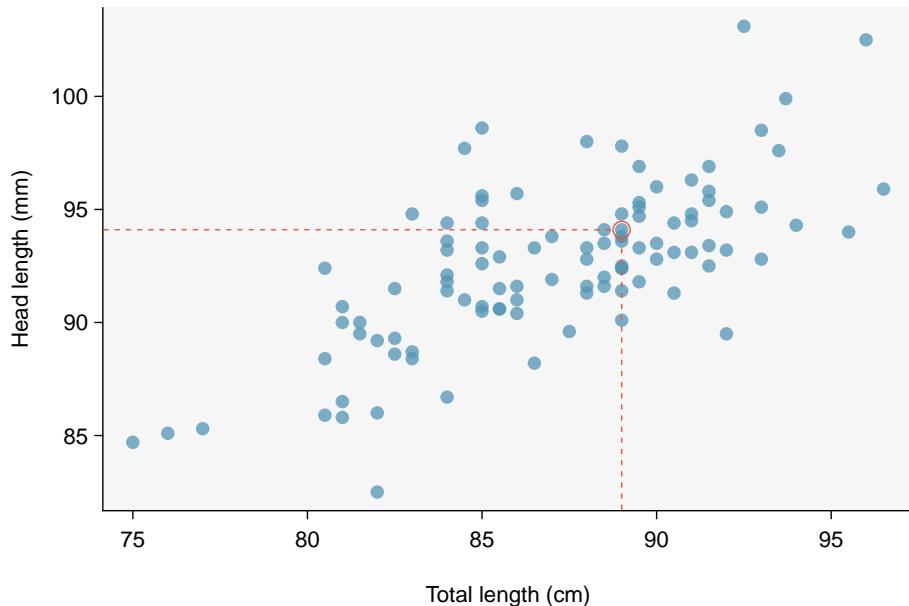
Figure 8.5: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1 mm and total length 89 cm is highlighted.

The head and total length variables are associated: possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length, x, to explain or predict a possum's head length, y. When we use x to predict y, we usually call x the **explanatory variable** or predictor variable, and we call y the **response variable**. We could fit the linear relationship by eye, as in Figure 8.6. The equation for this line is

$$\hat{y} = 41 + 0.59x$$

A "hat" on y is used to signify that this is a predicted value, not an observed value. We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

$$\hat{y} = 41 + 0.59(80)$$
$$= 88.2$$

The value $\hat{y}$ may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. The value $\hat{y}$ is also a prediction: absent further information about an 80 cm possum, this is our best prediction for a the head length of a single 80 cm possum.

8.1.3 Residuals

Residuals are the leftover variation in the response variable after fitting a model. Each observation will have a residual, and three of the residuals for the linear model we fit for the `possum` data are shown in Figure 8.6. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.
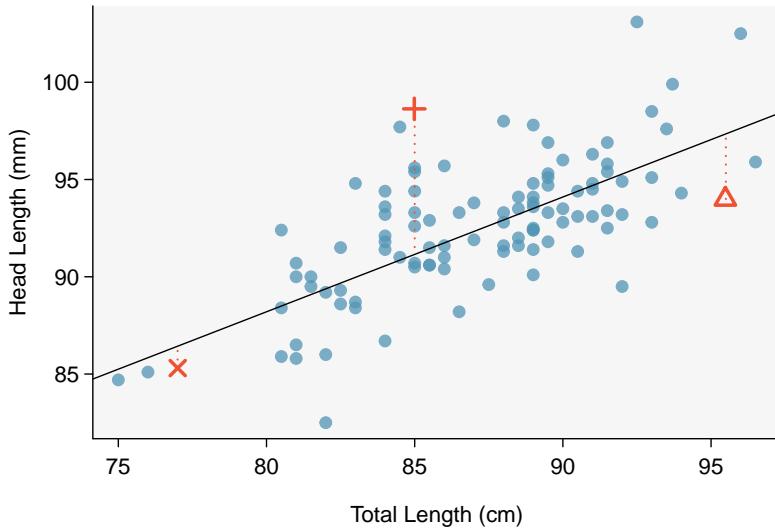
Figure 8.6: A reasonable linear model was fit to represent the relationship between head length and total length.

Let's look closer at the three residuals featured in Figure 8.6. The observation marked by an "×" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by "△" has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for "△" is larger than that of "×" because $|-4|$ is larger than $|-1|$.

RESIDUAL: DIFFERENCE BETWEEN OBSERVED AND EXPECTED

The residual for a particular observation $(x,\ y)$ is the difference between the observed response and the response we would predict based on the model:

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= y - \hat{y}$$

We typically identify $\hat{y}$ by plugging x into the model.

EXAMPLE 8.1

The linear fit shown in Figure 8.6 is given as $\hat{y} = 41 + 0.59x$. Based on this line, compute and interpret the residual of the observation (77.0, 85.3). This observation is denoted by "×" on the plot. Recall that x is the total length measured in cm and y is head length measured in mm.

——————

We first compute the predicted value based on the model:

$$\hat{y} = 41 + 0.59x$$
$$= 41 + 0.59(77.0)$$
$$= 86.4$$

Next we compute the difference of the actual head length and the predicted head length:

$$residual = y - \hat{y}$$
$$= 85.3 - 86.4$$
$$= -1.1$$

The residual for this point is -1.1 mm, which is very close to the visual estimate of -1 mm. For this particular possum with total length of 77 cm, the model's prediction for its head length was 1.1 mm *too high.*

GUIDED PRACTICE 8.2

If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

GUIDED PRACTICE 8.3

Compute the residual for the observation $(95.5, 94.0)$, denoted by "△" in the figure, using the linear model: $\hat{y} = 41 + 0.59x$.[2]

Residuals are helpful in evaluating how well a linear model fits a data set. We often display the residuals in a **residual plot** such as the one shown in Figure 8.7. Here, the residuals are calculated for each x value, and plotted versus x. For instance, the point $(85.0, 98.6)$ had a residual of 7.45, so in the residual plot it is placed at $(85.0, 7.45)$. Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal.

From the residual plot, we can better estimate the **standard deviation of the residuals**, often denoted by the letter s. The standard deviation of the residuals tells us typical size of the residuals. As such, it is a measure of the typical deviation between the y values and the model predictions. In other words, it tells us the typical prediction error using the model.[3]

————————————

[1]If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative.

[2]First compute the predicted value based on the model, then compute the residual.

$$\hat{y} = 41 + 0.59x = 41 + 0.59(95.50) = 97.3$$

$$residual = y - \hat{y} = 94.0 - 97.3 = -3.3$$

The residual is -3.3, so the model *overpredicted* the head length for this possum by 3.3 mm.

[3]The standard deviation of the residuals is calculated as: $s = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$.

EXAMPLE 8.4

Estimate the standard deviation of the residuals for predicting head length from total length using the line: $\hat{y} = 41 + 0.59x$ using Figure 8.7. Also, interpret the quantity in context.

To estimate this graphically, we use the residual plot. The approximate 68, 95 rule for standard deviations applies. Approximately 2/3 of the points are within $\pm$ 2.5 and approximately 95% of the points are within $\pm$ 5, so 2.5 is a good estimate for the standard deviation of the residuals. The typical error when predicting head length using this model is about 2.5 mm.
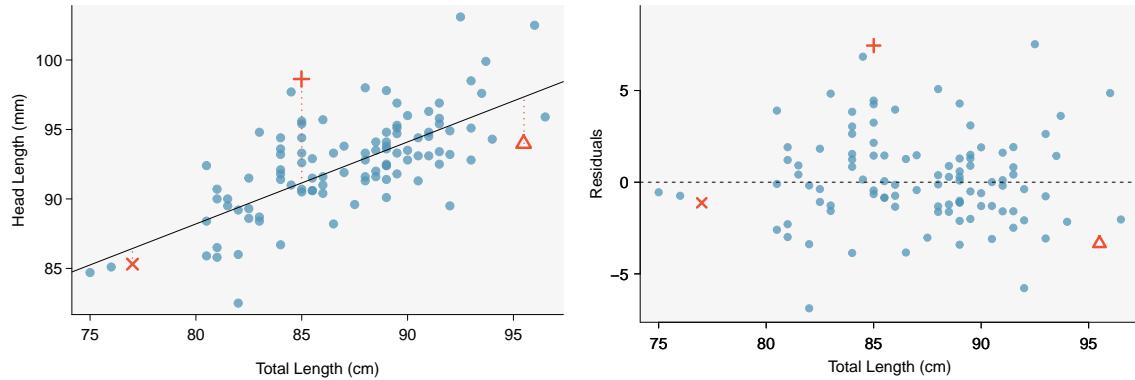
Figure 8.7: Left: Scatterplot of head length versus total length for 104 brushtail possums. Three particular points have been highlighted. Right: Residual plot for the model shown in left panel.

STANDARD DEVIATION OF THE RESIDUALS

The standard deviation of the residuals, often denoted by the letter s, tells us the typical error in the predictions using the regression model. It can be estimated from a residual plot.

EXAMPLE 8.5

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 8.8 shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

———————

In the first data set (first column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0.

In the second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is statistically significant evidence that the slope parameter is different from zero. The slope of the sample regression line is not zero, but we might wonder if this could be due to random variation. We will address this sort of scenario in Section 8.3.
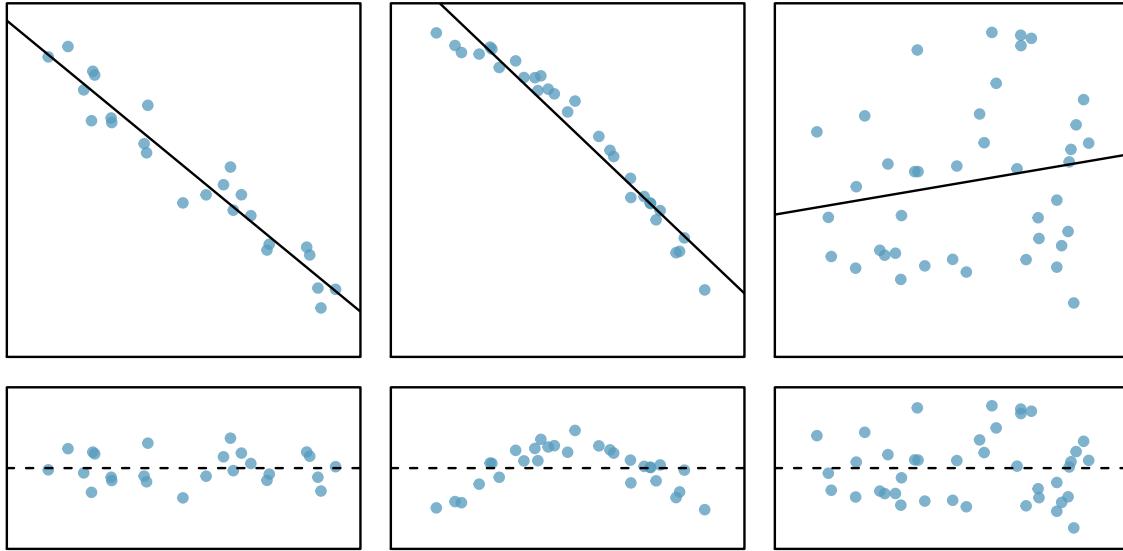
Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

8.1.4 Describing linear relationships with correlation

When a linear relationship exists between two variables, we can quantify the strength and direction of the linear relation with the correlation coefficient, or just **correlation** for short. Figure 8.9 shows eight plots and their corresponding correlations.

$r = 0.33$ $r = 0.69$ $r = 0.98$ $r = 1.00$

$r = -0.08$ $r = -0.64$ $r = -0.92$ $r = -1.00$
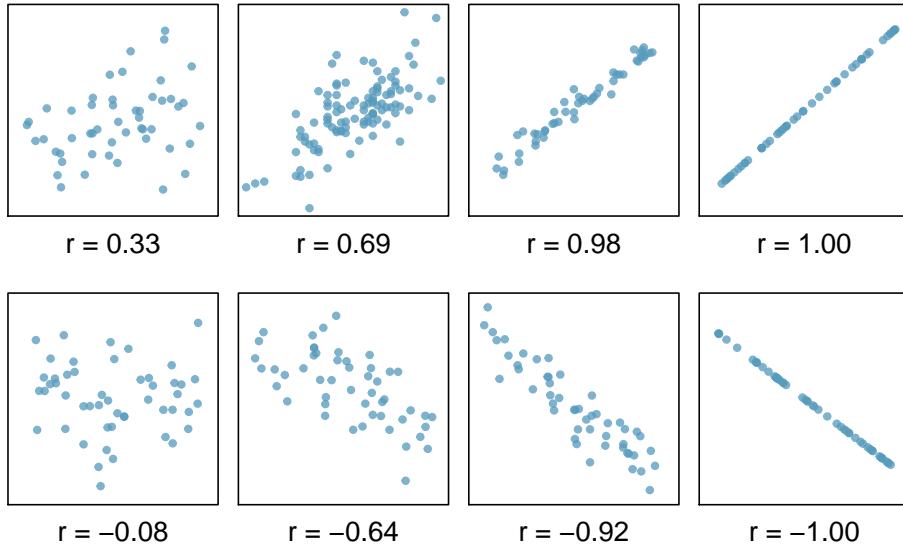
Figure 8.9: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

Only when the relationship is perfectly linear is the correlation either -1 or 1. If the linear relationship is strong and positive, the correlation will be near $+1$. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

CORRELATION MEASURES THE STRENGTH OF A LINEAR RELATIONSHIP

Correlation, which always takes values between -1 and 1, describes the direction and strength of the linear relationship between two numerical variables. The strength can be strong, moderate, or weak.

We compute the correlation using a formula, just as we did with the sample mean and standard deviation. Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) using the formula

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

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable. This formula is rather complex, and we generally perform the calculations on a computer or calculator. We can note, though, that the computation involves taking, for each point, the product of the Z-scores that correspond to the x and y values.

EXAMPLE 8.6

Take a look at Figure 8.6 on page 433. How would the correlation between head length and total body length of possums change if head length were measured in cm rather than mm? What if head length were measured in inches rather than mm?

Here, changing the units of y corresponds to multiplying all the y values by a certain number. This would change the mean and the standard deviation of y, but it would not change the correlation. To see this, imagine dividing every number on the vertical axis by 10. The units of y are now in cm rather than in mm, but the graph has remain exactly the same. The units of y have changed, by the relative distance of the y values about the mean are the same; that is, the Z-scores corresponding to the y values have remained the same.

CHANGING UNITS OF X AND Y DOES NOT AFFECT THE CORRELATION

The correlation, r, between two variables is not dependent upon the units in which the variables are recorded. Correlation itself has no units.

Correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 8.10.

r = −0.23   r = 0.31   r = 0.50
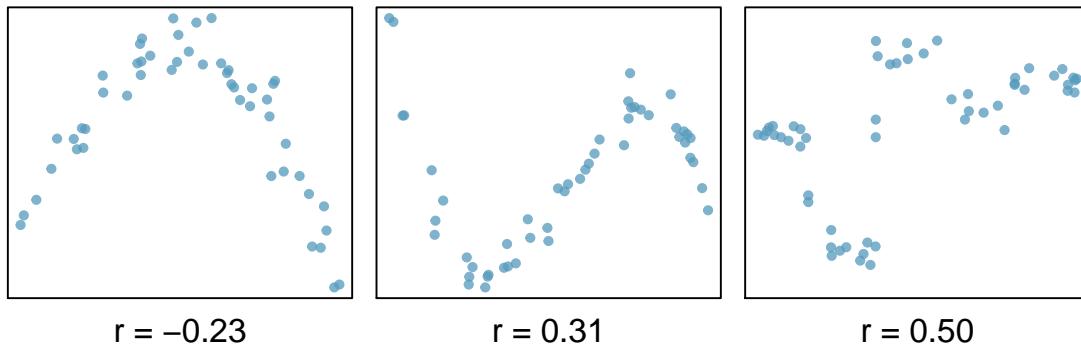
Figure 8.10: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

GUIDED PRACTICE 8.7

It appears no straight line would fit any of the datasets represented in Figure 8.10. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

---

[4]We'll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reflect overall trends in the data.

EXAMPLE 8.8

Consider the four scatterplots in Figure 8.11. In which scatterplot is the correlation between x and y the strongest?

(E)

All four data sets have the exact same correlation of $r = 0.816$ as well as the same equation for the best fit line! This group of four graphs, known as Anscombe's Quartet, remind us that knowing the value of the correlation does not tell us what the corresponding scatterplot looks like. It is always important to first graph the data. Investigate Anscombe's Quartet in Desmos: https://www.desmos.com/calculator/paknt6oneh.
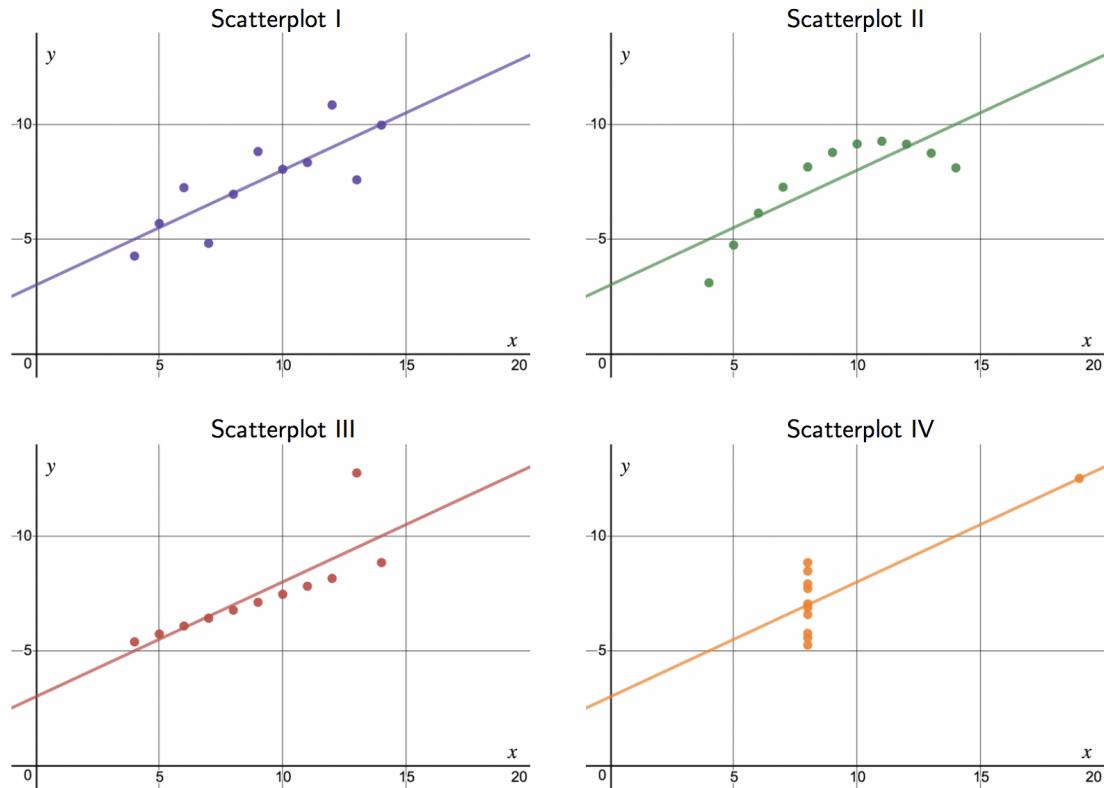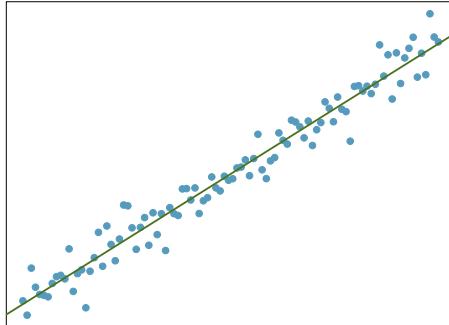
Figure 8.11: Four scatterplots from Desmos with best fit line drawn in.

Section summary
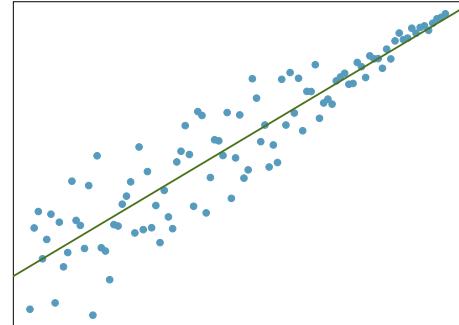
- In Chapter 2 we introduced a bivariate display called a **scatterplot**, which shows the relationship between two numerical variables. When we use x to predict y, we call x the **explanatory variable** or predictor variable, and we call y the **response variable**.

- A linear model for bivariate numerical data can be useful for prediction when the association between the variables follows a constant, linear trend. Linear models should not be used if the trend between the variables is curved.

- When we write a linear model, we use $\hat{y}$ to indicate that it is the model or the prediction. The value $\hat{y}$ can be understood as a **prediction** for y based on a given x, or as an **average** of the y values for a given x.

- The **residual** is the **error** between the true value and the modeled value, computed as $y - \hat{y}$. The order of the difference matters, and the sign of the residual will tell us if the model overpredicted or underpredicted a particular data point.

- The symbol s in a linear model is used to denote the standard deviation of the residuals, and it measures the typical prediction error by the model.

- A **residual plot** is a scatterplot with the residuals on the vertical axis. The residuals are often plotted against x on the horizontal axis, but they can also be plotted against y, $\hat{y}$, or other variables. Two important uses of a residual plot are the following.

  - Residual plots help us see patterns in the data that may not have been apparent in the scatterplot.

  - The standard deviation of the residuals is easier to estimate from a residual plot than from the original scatterplot.

- **Correlation**, denoted with the letter r, measures the strength and direction of a linear relationship. The following are some important facts about correlation.

  - The value of r is always between -1 and 1, inclusive, with an $r = -1$ indicating a perfect negative relationship (points fall exactly along a line that has negative slope) and an $r = 1$ indicating a perfect positive relationship (points fall exactly along a line that has positive slope).

  - An $r = 0$ indicates no *linear* association between the variables, though there may well exist a quadratic or other type of association.

  - Just like Z-scores, the correlation has no units. Changing the units in which x or y are measured does not affect the correlation.

  - Correlation is sensitive to outliers. Adding or removing a single point can have a big effect on the correlation.

  - As we learned previously, correlation is not causation. Even a very strong correlation cannot prove causation; only a well-designed, controlled, randomized experiment can prove causation.

Exercises

8.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.
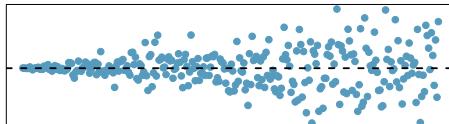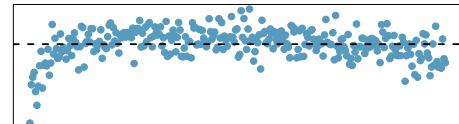
(a)   (b)

8.2 Trends in the residuals. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.
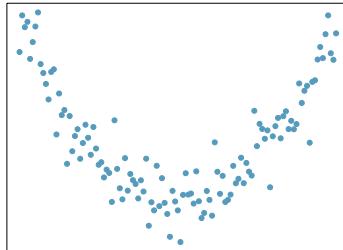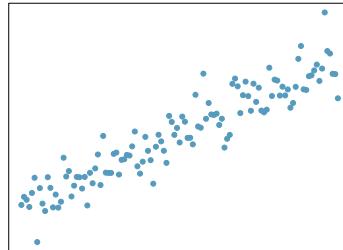
(a)   (b)
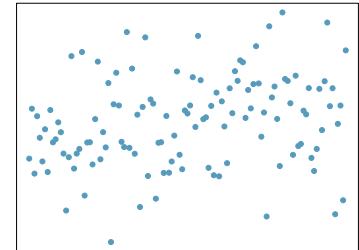
8.3 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.
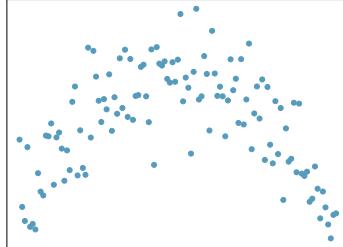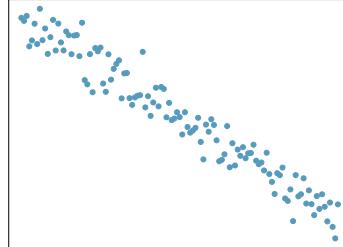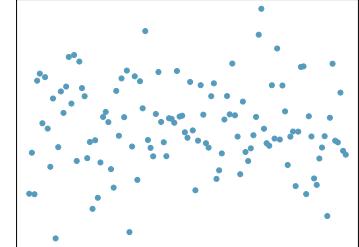
(a)   (b)   (c)
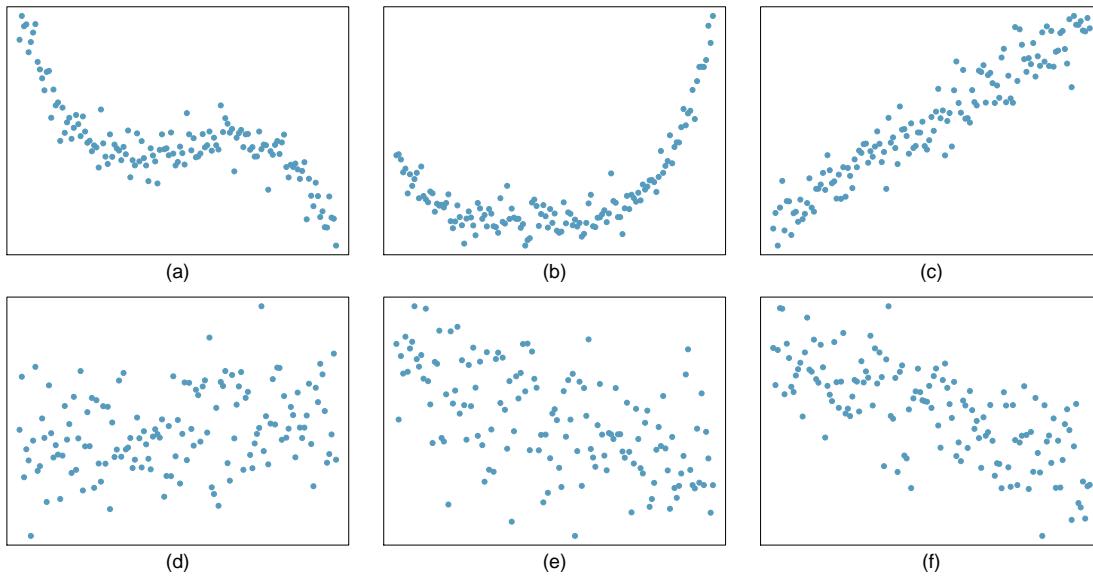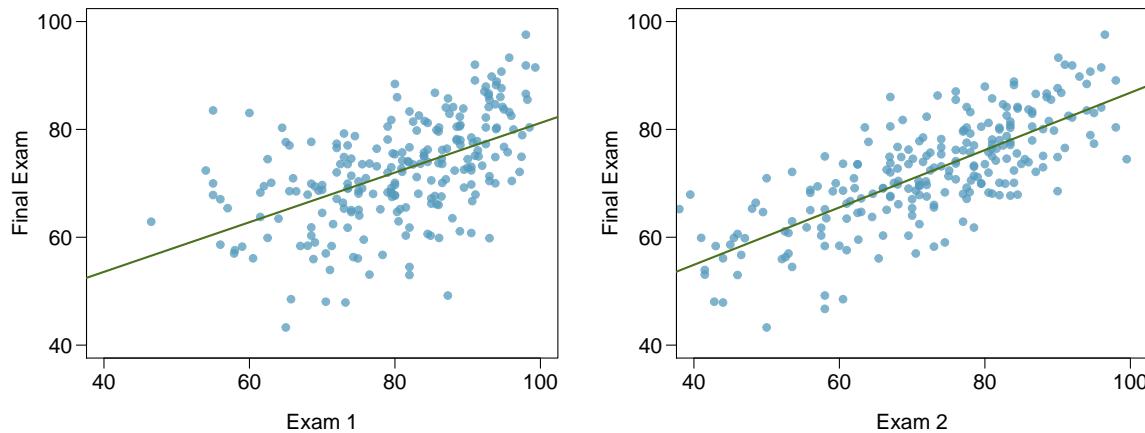
(d)   (e)   (f)

8.4   Identify relationships, Part II. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

(a)   (b)   (c)

(d)   (e)   (f)
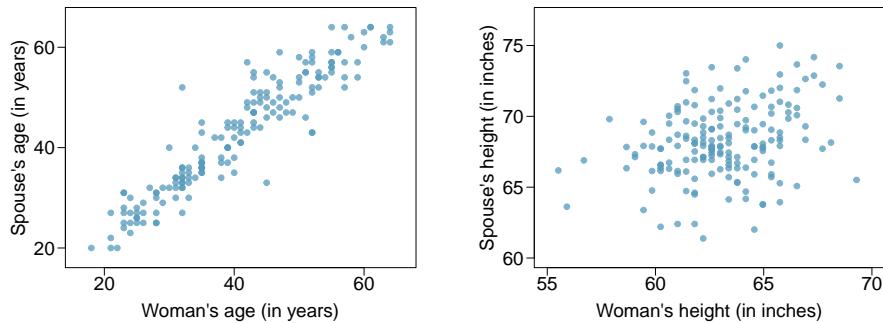
8.5   Exams and grades. The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

(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?
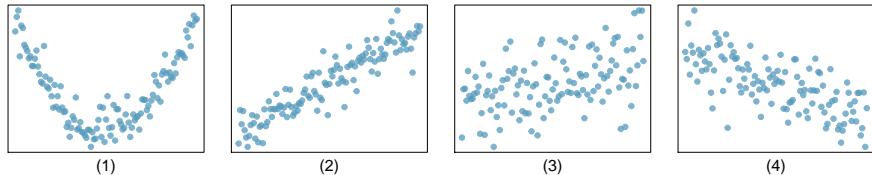
8.6 Spouses, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married women in Britain, recording the age (in years) and heights (converted here to inches) of the women and their spouses.[5] The scatterplot on the left shows the spouse's age plotted against the woman's age, and the plot on the right shows spouse's height plotted against the woman's height.

(a) Describe the relationship between the ages of women in the sample and their spouses' ages.
(b) Describe the relationship between the heights of women in the sample and their spouses' heights.
(c) Which plot shows a stronger correlation? Explain your reasoning.
(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between heights of women in the sample and their spouses' heights?
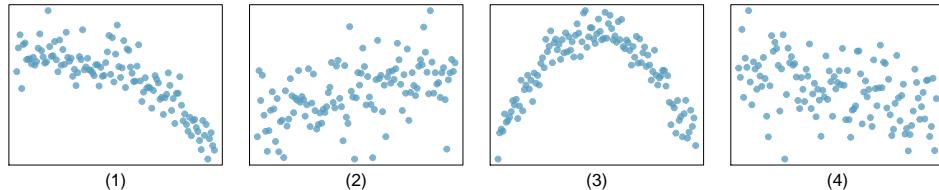
8.7 Match the correlation, Part I. Match each correlation to the corresponding scatterplot.

(a) $r = -0.7$
(b) $r = 0.45$
(c) $r = 0.06$
(d) $r = 0.92$
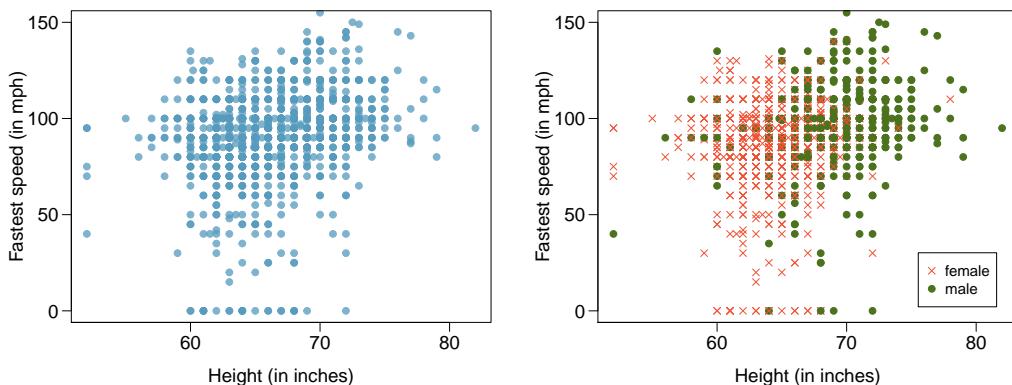
8.8 Match the correlation, Part II. Match each correlation to the corresponding scatterplot.

(a) $r = 0.49$
(b) $r = -0.48$
(c) $r = -0.03$
(d) $r = -0.85$

8.9 Speed and height. 1,302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.
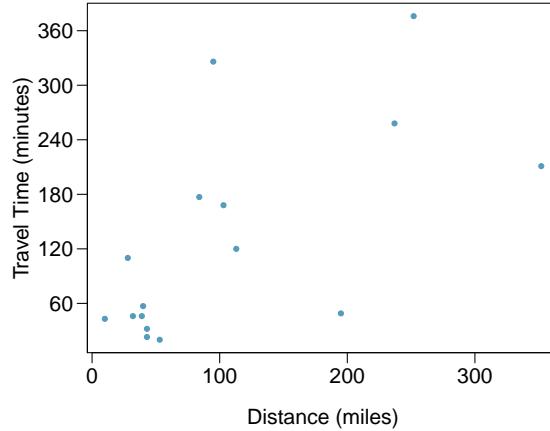
(a) Describe the relationship between height and fastest speed.
(b) Why do you think these variables are positively associated?
(c) What role does gender play in the relationship between height and fastest driving speed?

---

[5]D.J. Hand. *A handbook of small data sets.* Chapman & Hall/CRC, 1994.

8.10 Guess the correlation. Eduardo and Rosie are both collecting data on number of rainy days in a year and the total rainfall for the year. Eduardo records rainfall in inches and Rosie in centimeters. How will their correlation coefficients compare?
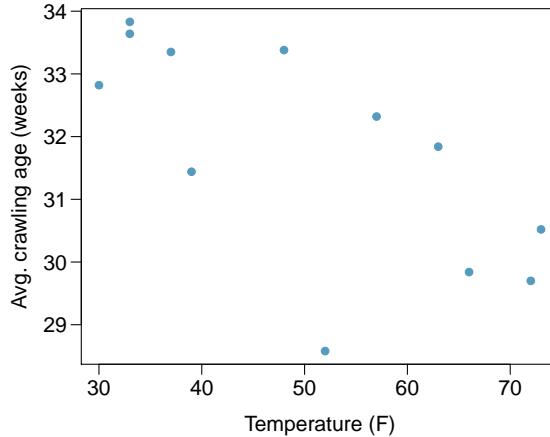
8.11 The Coast Starlight, Part I. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

(a) Describe the relationship between distance and travel time.
(b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?
(c) Correlation between travel time (in miles) and distance (in minutes) is $r = 0.636$. What is the correlation between travel time (in kilometers) and distance (in hours)?

8.12 Crawling babies, Part I. A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months.[6] Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (°F) and age is measured in weeks.
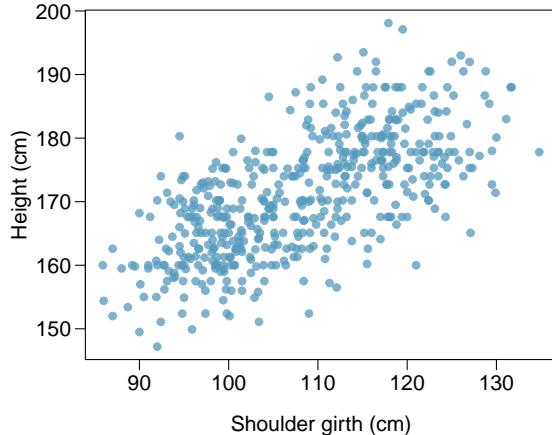
(a) Describe the relationship between temperature and crawling age.
(b) How would the relationship change if temperature was measured in degrees Celsius (°C) and age was measured in months?
(c) The correlation between temperature in °F and age in weeks was $r = -0.70$. If we converted the temperature to °C and age to months, what would the correlation be?

[6]J.B. Benson. "Season of birth and onset of locomotion: Theoretical and methodological implications". In: *Infant behavior and development* 16.1 (1993), pp. 69–81. ISSN: 0163-6383.
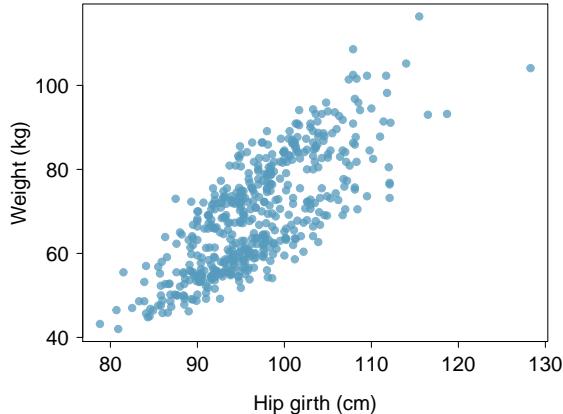
8.13   Body measurements, Part I. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.[7] The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

8.14   Body measurements, Part II. The scatterplot below shows the relationship between weight measured in kilograms and hip girth measured in centimeters from the data described in Exercise 8.13.

(a) Describe the relationship between hip girth and weight.

(b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?

8.15   Correlation, Part I. What would be the correlation between the ages of a set of women and their spouses if the set of women always married someone who was

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

8.16   Correlation, Part II. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

(a) $5,000 more than women?

(b) 25% more than women?

(c) 15% less than women?

---

[7]G. Heinz et al. "Exploring relationships in body dimensions". In: *Journal of Statistics Education* 11.2 (2003).

8.2   Fitting a line by least squares regression

In this section, we answer the following questions:

- How well can we predict financial aid based on family income for a particular college?

- How does one find, interpret, and apply the least squares regression line?

- How do we measure the fit of a model and compare different models to each other?

- Why do models sometimes make predictions that are ridiculous or impossible?

Learning objectives

1. Calculate the slope and y-intercept of the least squares regression line using the relevant summary statistics. Interpret these quantities in context.

2. Understand why the least squares regression line is called the least squares regression line.

3. Interpret the explained variance R^2.

4. Understand the concept of extrapolation and why it is dangerous.

5. Identify outliers and influential points in a scatterplot.

8.2.1   An objective measure for finding the best line

Fitting linear models by eye is open to criticism since it is based on an individual preference. In this section, we use *least squares regression* as a more rigorous approach.

This section considers family income and gift aid data from a random sample of fifty students in the freshman class of Elmhurst College in Illinois.[8] Gift aid is financial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 8.12 along with two linear fits. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university.

We begin by thinking about what we mean by "best". Mathematically, we want a line that has small residuals. Perhaps our criterion could minimize the sum of the residual magnitudes:

$$|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \cdots + |y_n - \hat{y}_n|$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 8.12 demonstrates this fit can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the squared residuals:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

---

[8]These data were sampled from a table of data for all freshmen from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435
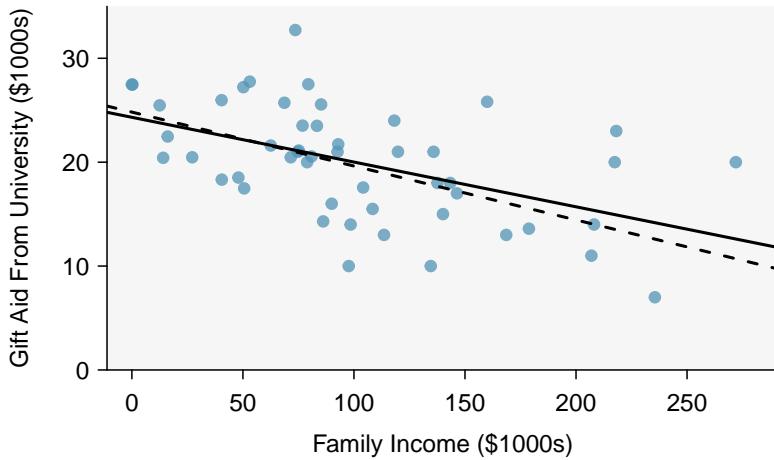
Figure 8.12: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

The line that minimizes the sum of the squared residuals is represented as the solid line in Figure 8.12. This is commonly called the **least squares line**.

Both lines seem reasonable, so why do data scientists prefer the least squares regression line? One reason is that it is easier to compute by hand and in most statistical software. Another, and more compelling, reason is that in many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.

In Figure 8.13, we imagine the squared error about a line as actual squares. The least squares regression line minimizes the sum of the *areas* of these squared errors. In the figure, the sum of the squared error is $4 + 1 + 1 = 6$. There is no other line about which the sum of the squared error will be smaller.
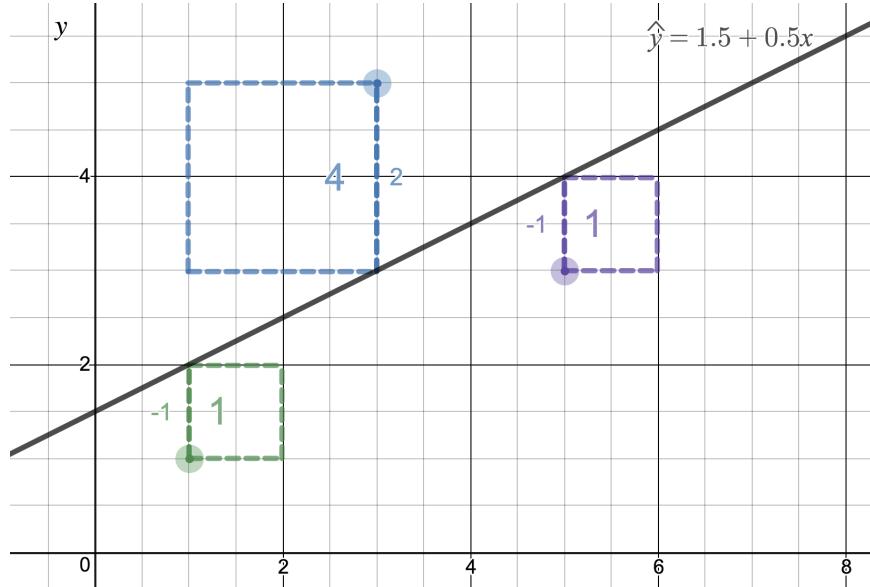
Figure 8.13: A visualization of least squares regression using Desmos. Try out this and other interactive Desmos activities at openintro.org/ahss/desmos.

8.2.2 Finding the least squares line

For the Elmhurst College data, we could fit a least squares regression line for predicting gift aid based on a student's family income and write the equation as:

$$\widehat{aid} = a + b \times family\_income$$

Here a is the y-intercept of the least squares regression line and b is the slope of the least squares regression line. a and b are both statistics that can be calculated from the data. In the next section we will consider the corresponding parameters that they statistics attempt to estimate.

We can enter all of the data into a statistical software package and easily find the values of a and b. However, we can also calculate these values by hand, using only the summary statistics.

- The slope of the least squares line is given by

$$b = r \frac{s_y}{s_x}$$

  where r is the correlation between the variables x and y, and s_x and s_y are the sample standard deviations of x, the explanatory variable, and y, the response variable.

- The point of averages $(\bar{x}, \bar{y})$ is always on the least squares line. Plugging this point in for x and y in the least squares equation and solving for a gives

$$\bar{y} = a + b\bar{x} \qquad\qquad a = \bar{y} - b\bar{x}$$

FINDING THE SLOPE AND INTERCEPT OF THE LEAST SQUARES REGRESSION LINE

The least squares regression line for predicting y based on x can be written as: $\hat{y} = a + bx$.

$$b = r \frac{s_y}{s_x} \qquad \bar{y} = a + b\bar{x}$$

We first find b, the slope, and then we solve for a, the y-intercept.

GUIDED PRACTICE 8.9

Figure 8.14 shows the sample means for the family income and gift aid as \$101,800 and \$19,940, respectively. Plot the point $(101.8, 19.94)$ on Figure 8.12 to verify it falls on the least squares line (the solid line).[9]

	family income, in \$1000s ("x")	gift aid, in \$1000s ("y")
mean	$\bar{x} = 101.8$	$\bar{y} = 19.94$
sd	$s_x = 63.2$	$s_y = 5.46$
		$r = -0.499$

Figure 8.14: Summary statistics for family income and gift aid.

---

[9]If you need help finding this location, draw a straight line up from the x-value of 100 (or thereabout). Then draw a horizontal line at 20 (or thereabout). These lines should intersect on the least squares line.

EXAMPLE 8.10

Using the summary statistics in Figure 8.14, find the equation of the least squares regression line for predicting gift aid based on family income.

(E)

$$b = r\frac{s_y}{s_x} = (-0.499)\frac{5.46}{63.2} = -0.0431$$
$$a = \bar{y} - b\bar{x} = 19.94 - (-0.0431)(101.8) = 24.3$$

$$\hat{y} = 24.3 - 0.0431x \qquad \text{or} \qquad \widehat{aid} = 24.3 - 0.0431 \times family\_income$$

EXAMPLE 8.11

Say we wanted to predict a student's family income based on the amount of gift aid that they received. Would this least squares regression line be the following?

$$aid = 24.3 - 0.0431 \times \widehat{family\_income}$$

(E)

No. The equation we found was for predicting aid, not for predicting family income. We would have to calculate a new regression line, letting y be *family_income* and x be *aid*. This would give us:

$$b = r\frac{s_y}{s_x} = (-0.499)\frac{63.2}{5.46} = -5.776$$
$$a = \bar{y} - b\bar{x} = 19.94 - (-5.776)(101.8) = 607.9$$

$$\hat{y} = 607.3 - 5.776x \qquad \text{or} \qquad \widehat{family\_income} = 607.3 - 5.776 \times aid$$

We mentioned earlier that a computer is usually used to compute the least squares line. A summary table based on computer output is shown in Figure 8.15 for the Elmhurst College data. The first column of numbers provides estimates for b_0 and b_1, respectively. Compare these to the result from Example 8.2.2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.15: Summary of least squares fit for the Elmhurst College data. Compare the parameter estimates in the first column to the results of Guided Practice 8.2.2.

EXAMPLE 8.12

Examine the second, third, and fourth columns in Figure 8.15. Can you guess what they represent?

(E)

We'll look at the second row, which corresponds to the slope. The first column, Estimate = -0.0431, tells us our best estimate for the slope of the population regression line. We call this point estimate b. The second column, Std. Error = 0.0108, is the standard error of this point estimate. The third column, t value = -3.98, is the T test statistic for the null hypothesis that the slope of the population regression line = 0. The last column, Pr(>|t|) = 0.0002, is the p-value for this two-sided T-test. We will get into more of these details in Section 8.3.

EXAMPLE 8.13

Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have found to calculate her financial aid from the university?

(E)

No. Using the equation will provide a prediction or estimate. However, as we see in the scatterplot, there is a lot of variability around the line. While the linear equation is good at capturing the trend in the data, there will be significant error in predicting an individual student's aid. Additionally, the data all come from one freshman class, and the way aid is determined by the university may change from year to year.

8.2.3   Interpreting the coefficients of a regression line

Interpreting the coefficients in a regression model is often one of the most important steps in the analysis.

EXAMPLE 8.14

The slope for the Elmhurst College data for predicting gift aid based on family income was calculated as -0.0431. Intepret this quantity in the context of the problem.

(E)

You might recall from an algebra course that slope is change in y over change in x. Here, both x and y are in thousands of dollars. So if x is one unit or one thousand dollars higher, the line will predict that y will change by 0.0431 thousand dollars. In other words, for each additional thousand dollars of family income, *on average*, students receive 0.0431 thousand, or $43.10 *less* in gift aid. Note that a higher family income corresponds to less aid because the slope is negative.

EXAMPLE 8.15

The y-intercept for the Elmhurst College data for predicting gift aid based on family income was calculated as 24.3. Intepret this quantity in the context of the problem.

(E)

The intercept a describes the predicted value of y when $x = 0$. The *predicted* gift aid is 24.3 thousand dollars if a student's family has no income. The meaning of the intercept is relevant to this application since the family income for some students at Elmhurst is $0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero. Here, it would be acceptable to say that the *average* gift aid is 24.3 thousand dollars among students whose family have 0 dollars in income.

INTERPRETING COEFFICIENTS IN A LINEAR MODEL

- The slope, b, describes the *average* increase or decrease in the y variable if the explanatory variable x is one unit larger.

- The y-intercept, a, describes the predicted outcome of y if $x = 0$. The linear model must be valid all the way to $x = 0$ for this to make sense, which in many applications is not the case.

(G)

GUIDED PRACTICE 8.16

In the previous chapter, we encountered a data set that compared the price of new textbooks for UCLA courses at the UCLA Bookstore and on Amazon. We fit a linear model for predicting price at UCLA Bookstore from price on Amazon and we get:

$$\hat{y} = 1.86 + 1.03x$$

where x is the price on Amazon and y is the price at the UCLA bookstore. Interpret the coefficients in this model and discuss whether the interpretations make sense in this context.[10]

(G)

GUIDED PRACTICE 8.17

Can we conclude that if Amazon raises the price of a textbook by 1 dollar, the UCLA Bookstore will raise the price of the textbook by $1.03?[11]

EXERCISE CAUTION WHEN INTERPRETING COEFFICIENTS OF A LINEAR MODEL

- The slope tells us only the *average* change in y for each unit change in x; it does not tell us how much y might change based on a change in x for any particular *individual*. Moreover, in most cases, the slope cannot be interpreted in a causal way.
- When a value of $x = 0$ doesn't make sense in an application, then the interpretation of the y-intercept won't have any practical meaning.

8.2.4 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 6th, 2010 [12]

Linear models can be used to approximate the relationship between two 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.

---

[10]The y-intercept is 1.86 and the units of y are in dollars. This tells us that when a textbook costs 0 dollars on Amazon, the predicted price of the textbook at the UCLA Bookstore is 1.86 dollars. This does not make sense as Amazon does not sell any $0 textbooks. The slope is 1.03, with units (dollars)/(dollars). On average, for every extra dollar that a book costs on Amazon, it costs an extra 1.03 dollars at the UCLA Bookstore. This interpretation does make sense in this context.

[11]No. The slope describes the overall trend. This is observational data; a causal conclusion cannot be drawn. Remember, a causal relationship can only be concluded by a well-designed randomized, controlled experiment. Additionally, there may be large variation in the points about the line. The slope does not tell us how much y might change based on a change in x for a particular textbook.

[12]www.cc.com/video-clips/l4nkoq/

EXAMPLE 8.18

Use the model $\widehat{aid} = 24.3 - 0.0431 \times family\_income$ to estimate the aid of another freshman student whose family had income of $1 million.

Recall that the units of family income are in $1000s, so we want to calculate the aid for $family\_income = 1000$:

$$\widehat{aid} = 24.3 - 0.0431 \times family\_income$$
$$\widehat{aid} = 24.3 - 0.431(1000) = -18.8$$

The model predicts this student will have -$18,800 in aid (!). Elmhurst College cannot (or at least does not) require any students to pay extra on top of tuition to attend.

Using a model to predict y-values for x-values outside the domain of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

8.2.5   Using R^2 to describe the strength of a fit

We evaluated the strength of the linear relationship between two variables earlier using the correlation, r. However, it is more common to explain the fit of a model using R^2, called **R-squared** or the **explained variance**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.
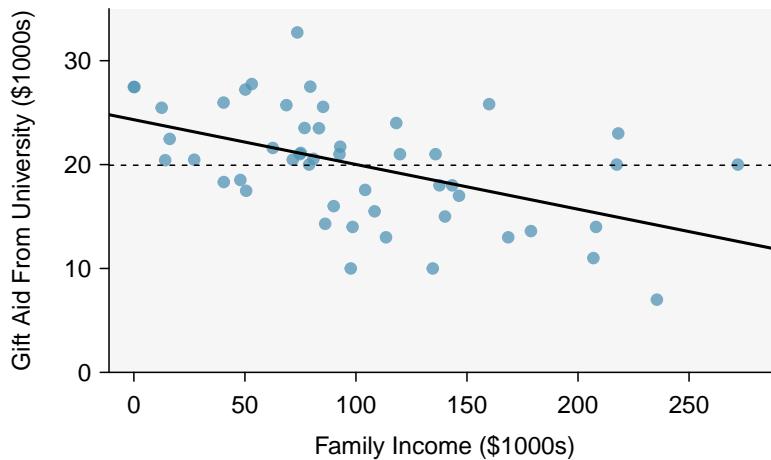
Figure 8.16: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line ($\hat{y}$) and the average line ($\bar{y}$).

We are interested in how well a model accounts for or explains the location of the y values. The R^2 of a linear model describes how much smaller the variance (in the y direction) about the regression line is than the variance about the horizontal line $\bar{y}$. For example, consider the Elmhurst College data, shown in Figure 8.16. The variance of the response variable, aid received, is $s^2_{aid} = 29.8$. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's family income. The variability in the residuals describes how much variation remains after using the model: $s^2_{RES} = 22.4$. We could say that the reduction in the variance was:

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29.8 - 22.4}{29.8} = \frac{7.5}{29.8} = 0.25$$

If we used the simple standard deviation of the residuals, this would be exactly R^2. However, the standard way of computing the standard deviation of the residuals is slightly more sophisticated.[13] To avoid any trouble, we can instead use a sum of squares method. If we call the sum of the squared errors about the regression line $SSRes$ and the sum of the squared errors about the mean SSM, we can define R^2 as follows:

$$R^2 = \frac{SSM - SSRes}{SSM} = 1 - \frac{SSRes}{SSM}$$

(a)

(b)
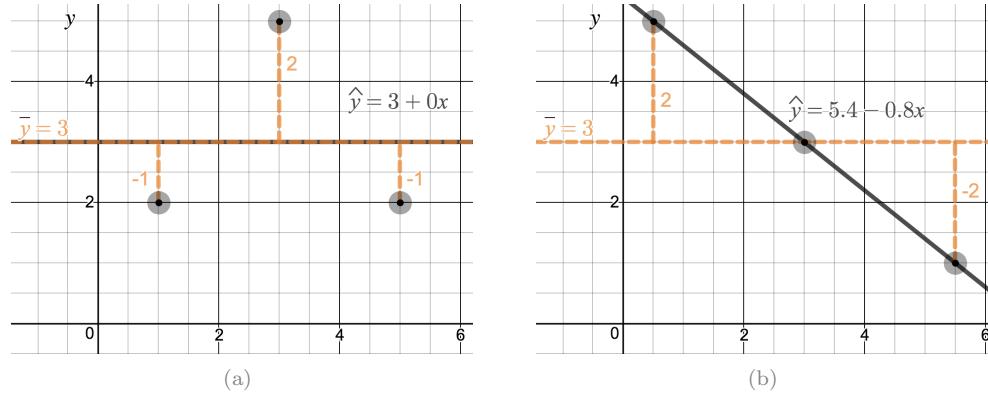
Figure 8.17: (a) The regression line is equivalent to $\bar{y}$; $R^2 = 0$. (b) The regression line passes through all of the points; $R^2 = 1$. Try out this and other interactive Desmos activities at openintro.org/ahss/desmos.

GUIDED PRACTICE 8.19

Using the formula for R^2, confirm that in Figure 8.17 (a), $R^2 = 0$ and that in Figure 8.17 (b), $R^2 = 1$.[14]

R^2 IS THE EXPLAINED VARIANCE

R^2 is always between 0 and 1, inclusive. It tells us the proportion of variation in the y values that is explained by a regression model. The higher the value of R^2, the better the model "explains" the response variable.

The value of R^2 is, in fact, equal to r^2, where r is the correlation. This means that $r = \pm\sqrt{R^2}$. Use this fact to answer the next two practice problems.

GUIDED PRACTICE 8.20

If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response variable is explained by the linear model?[15]

GUIDED PRACTICE 8.21

If a linear model has an R^2 or explained variance of 0.94, what is the correlation?[16]

---

[13]In computing the standard deviation of the residuals, we divide by $n - 2$ rather than by $n - 1$ to account for the $n - 2$ degrees of freedom.

[14](a) $SSRes = SSM = (-1)^2 + (2)^2 + (-1)^2 = 6$, so $R^2 = 1 - \frac{6}{6} = 0$. (b) $R^2 = 1 - \frac{0}{8}$.

[15]$R^2 = (-0.97)^2 = 0.94$ or 94%. 94% of the variation in y is explained by the linear model.

[16]We take the square root of R^2 and get 0.97, but we must be careful, because r could be 0.97 or -0.97. Without knowing the slope or seeing the scatterplot, we have no way of knowing if r is positive or negative.

8.2.6   Calculator/Desmos: linear correlation and regression

◆ TI-84: FINDING a, b, R^2, AND r FOR A LINEAR MODEL

Use STAT, CALC, LinReg(a + bx).

1. Choose STAT.
2. Right arrow to CALC.
3. Down arrow and choose 8:LinReg(a+bx).
   - Caution: choosing 4:LinReg(ax+b) will reverse a and b.
4. Let Xlist be L1 and Ylist be L2 (don't forget to enter the x and y values in L1 and L2 before doing this calculation).
5. Leave FreqList blank.
6. Leave Store RegEQ blank.
7. Choose Calculate and hit ENTER, which returns:
a	a, the y-intercept of the best fit line
b	b, the slope of the best fit line
r^2	R^2, the explained variance
r	r, the correlation coefficient

TI-83: Do steps 1-3, then enter the x list and y list separated by a comma, e.g. LinReg(a+bx) L1, L2, then hit ENTER.

WHAT TO DO IF R^2 AND r DO NOT SHOW UP ON A TI-83/84

If r^2 and r do now show up when doing STAT, CALC, LinReg, the *diagnostics* must be turned on. This only needs to be once and the diagnostics will remain on.

1. Hit 2ND 0 (i.e. CATALOG).
2. Scroll down until the arrow points at DiagnosticOn.
3. Hit ENTER and ENTER again. The screen should now say:

   DiagnosticOn
                       Done

WHAT TO DO IF A TI-83/84 RETURNS: ERR: DIM MISMATCH

This error means that the lists, generally L1 and L2, do not have the same length.

1. Choose 1:Quit.
2. Choose STAT, Edit and make sure that the lists have the same number of entries.

▶ **CASIO FX-9750GII: FINDING a, b, R^2, AND r FOR A LINEAR MODEL**

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).

2. Enter the x and y data into 2 separate lists, e.g. x values in List 1 and y values in List 2. Observation ordering should be the same in the two lists. For example, if $(5, 4)$ is the second observation, then the second value in the x list should be 5 and the second value in the y list should be 4.

3. Navigate to CALC (F2) and then SET (F6) to set the regression context.

   - To change the 2Var XList, navigate to it, select List (F1), and enter the proper list number. Similarly, set 2Var YList to the proper list.

4. Hit EXIT.

5. Select REG (F3), X (F1), and a+bx (F2), which returns:

a	a, the y-intercept of the best fit line
b	b, the slope of the best fit line
r	r, the correlation coefficient
r²	R^2, the explained variance
MSe	Mean squared error, which you can ignore

   If you select ax+b (F1), the a and b meanings will be reversed.

GUIDED PRACTICE 8.22

The data set loan50, introduced in Chapter 1, contains information on randomly sampled loans offered through Lending Club. A subset of the data matrix is shown in Figure 8.18. Use a calculator to find the equation of the least squares regression line for predicting loan amount from total income.[17]

	total_income	loan_amount
1	59000	22000
2	60000	6000
3	75000	25000
4	75000	6000
5	254000	25000
6	67000	6400
7	28800	3000
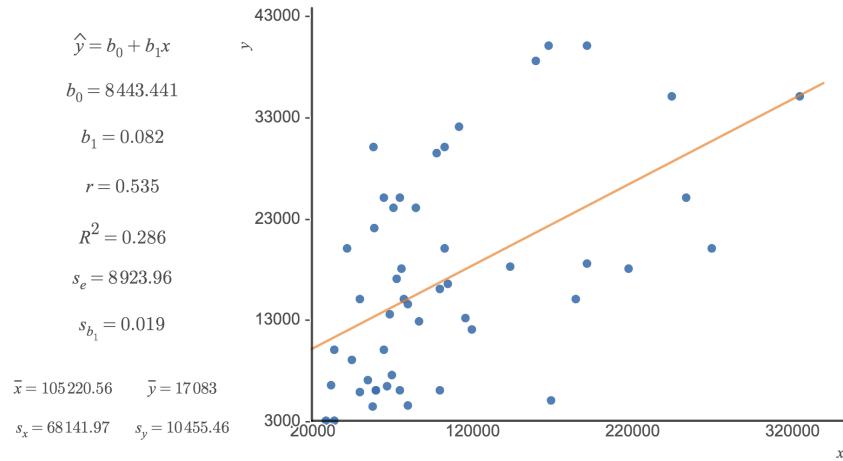
Figure 8.18: Sample of data from loan50.

---

[17]a = 11121 and b = 0.0043, therefore $\hat{y} = 11121 + 0.0043x$.

EXAMPLE 8.23

Use the full `loan50` data set (openintro.org/ahss/data) and this Desmos Calculator (openintro.org/ahss/desmos) to draw the scatterplot and find the equation of the least squares regression line for prediction loan amount (y) from total income (x).

———————

$\widehat{y} = b_0 + b_1 x$

$b_0 = 8\,443.441$

$b_1 = 0.082$

$r = 0.535$

$R^2 = 0.286$

$s_e = 8\,923.96$

$s_{b_1} = 0.019$

$\bar{x} = 105\,220.56 \qquad \bar{y} = 17\,083$

$s_x = 68\,141.97 \qquad s_y = 10\,455.46$

8.2.7 Types of outliers in linear regression

Outliers in regression are observations that fall far from the "cloud" of points. These points are especially important because they can have a strong influence on the least squares line.

EXAMPLE 8.24

There are six plots shown in Figure 8.19 along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify any obvious outliers and note how they influence the least squares line. Recall that an outlier is any point that doesn't appear to belong with the vast majority of the other points.

(1) There is one outlier far from the other points, though it only appears to slightly influence the line.

(2) There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.

(3) There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well.

(4) There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least squares line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

(5) There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.

(6) There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential.

Examine the residual plots in Figure 8.19. You will probably find that there is some trend in the main clouds of (3) and (4). In these cases, the outliers influenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

LEVERAGE

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage**.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in cases (3), (4), and (5) of Example 8.24 – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

DON'T IGNORE OUTLIERS WHEN FITTING A FINAL MODEL

If there are outliers in the data, they should not be removed or ignored without a good reason. Whatever final model is fit to the data would not be very helpful if it ignores the most exceptional cases.
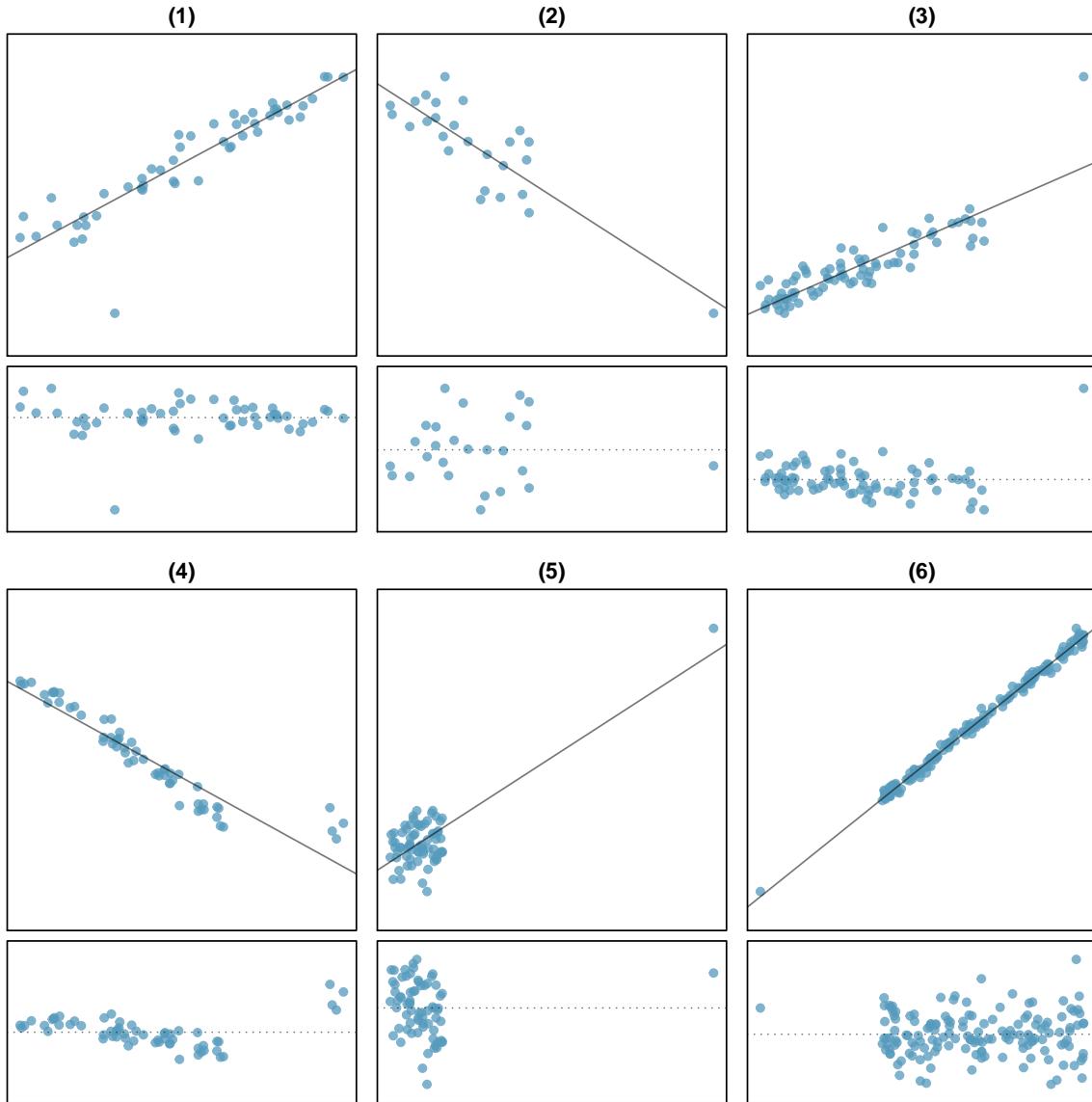

Figure 8.19: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

8.2.8 Categorical predictors with two levels (special topic)

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*). We'll consider eBay auctions for a video game, *Mario Kart* for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded.[18] Here we want to predict total price based on game condition, which takes values `used` and `new`. A plot of the auction data is shown in Figure 8.20.
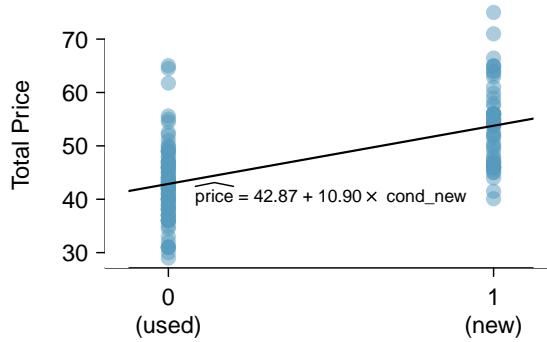

Figure 8.20: Total auction prices for the game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games with the least squares regression line shown.

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `cond_new`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{price} = \alpha + \beta \times \texttt{cond\_new}$$

The fitted model is summarized in Figure 8.21, and the model with its parameter estimates is given as

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

For categorical predictors with two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal with equal variance. Based on Figure 8.20, both of these conditions are reasonably satisfied.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	0.0000
cond_new	10.90	1.26	8.66	0.0000

Figure 8.21: Least squares regression summary for the Mario Kart data.

EXAMPLE 8.25

Interpret the two parameters estimated in the model for the price of *Mario Kart* in eBay auctions.

The intercept is the estimated price when `cond_new` takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.

INTERPRETING MODEL ESTIMATES FOR CATEGORICAL PREDICTORS.

The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

---

[18]These data were collected in Fall 2009 and may be found at openintro.org/stat.

Section summary

- We define the *best fit line* as the line that minimizes the sum of the squared residuals (errors) about the line. That is, we find the line that minimizes $(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = \sum (y_i - \hat{y}_i)^2$. We call this line the **least squares regression line**.

- We write the least squares regression line in the form: $\hat{y} = a + bx$, and we can calculate a and b based on the summary statistics as follows:

$$b = r \frac{s_y}{s_x} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}.$$

- *Interpreting* the **slope** and **y-intercept** of a linear model

    - The slope, b, describes the *average* increase or decrease in the y variable if the explanatory variable x is one unit larger.

    - The y-intercept, a, describes the average or predicted outcome of y if $x = 0$. The linear model must be valid all the way to $x = 0$ for this to make sense, which in many applications is not the case.

- Two important considerations about the regression line

    - The regression line provides *estimates* or *predictions*, not actual values. It is important to know how large s, the standard deviation of the residuals, is in order to know about how much error to expect in these predictions.

    - The regression line estimates are only reasonable within the domain of the data. Predicting y for x values that are outside the domain, known as **extrapolation**, is unreliable and may produce ridiculous results.

- Using R^2 to assess the fit of the model

    - R^2, called **R-squared** or the **explained variance**, is a measure of how well the model explains or fits the data. R^2 is always between 0 and 1, inclusive, or between 0% and 100%, inclusive. The higher the value of R^2, the better the model "fits" the data.

    - The R^2 for a linear model describes the *proportion of variation* in the y variable that is *explained by* the regression line.

    - R^2 applies to any type of model, not just a linear model, and can be used to compare the fit among various models.

    - The correlation $r = -\sqrt{R^2}$ or $r = \sqrt{R^2}$. The value of R^2 is always positive and cannot tell us the *direction* of the association. If finding r based on R^2, make sure to use either the scatterplot or the slope of the regression line to determine the *sign* of r.

- When a residual plot of the data appears as a random cloud of points, a linear model is generally appropriate. If a residual plot of the data has any type of pattern or curvature, such as a ∪-shape, a linear model is not appropriate.

- **Outliers** in regression are observations that fall far from the "cloud" of points.

- An **influential point** is a point that has a big effect or pull on the slope of the regression line. Points that are outliers in the x direction will have more pull on the slope of the regression line and are more likely to be influential points.

Exercises

8.17   Units of regression. Consider a regression predicting weight (kg) from height (cm) for a sample of adult males. What are the units of the correlation coefficient, the intercept, and the slope?
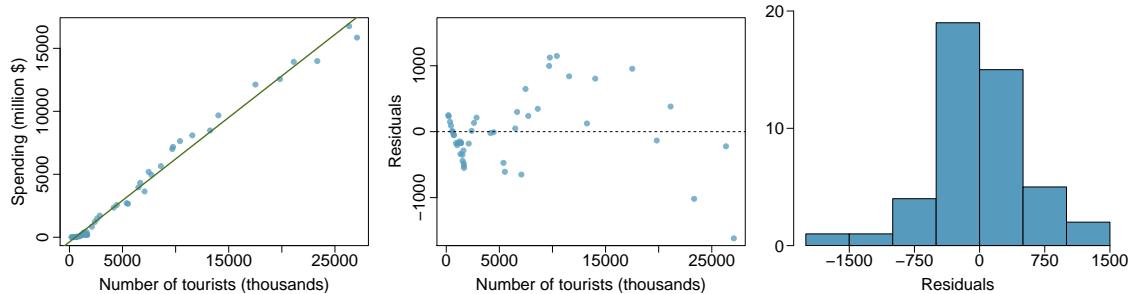
8.18   Which is higher? Determine if I or II is higher or if they are equal. Explain your reasoning. For a regression line, the uncertainty associated with the slope estimate, b_1, is higher when

   I. there is a lot of scatter around the regression line or

  II. there is very little scatter around the regression line

8.19   Over-under, Part I. Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under estimate the shelf-life of the apple? Explain your reasoning.

8.20   Over-under, Part II. Suppose we fit a regression line to predict the number of incidents of skin cancer per 1,000 people from the number of sunny days in a year. For a particular year, we predict the incidence of skin cancer to be 1.5 per 1,000 people, and the residual for this year is 0.5. Did we over or under estimate the incidence of skin cancer? Explain your reasoning.
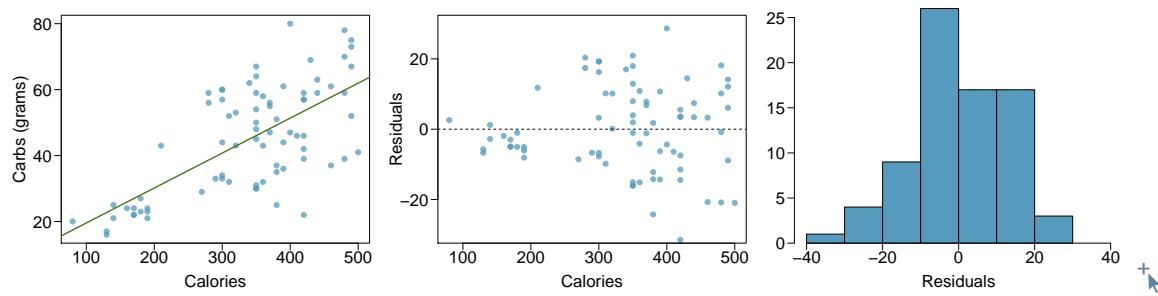
8.21   Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.[19] Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.

(a) Describe the relationship between number of tourists and spending.

(b) What are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

---

[19]Association of Turkish Travel Agencies, Foreign Visitors Figure & Tourist Spendings By Years.

8.22   Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.[20] Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.


(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

(b) In this scenario, what are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do these data meet the conditions required for fitting a least squares line?

8.23   The Coast Starlight, Part II. Exercise 8.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.

(d) 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.

(e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

8.24   Body measurements, Part III. Exercise 8.13 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.
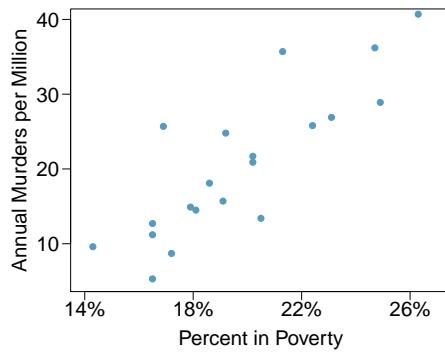
(a) Write the equation of the regression line for predicting height.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

8.25 Murders and poverty, Part I. ▶ The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $\quad R^2 = 70.52\%$ $\quad R^2_{adj} = 68.89\%$
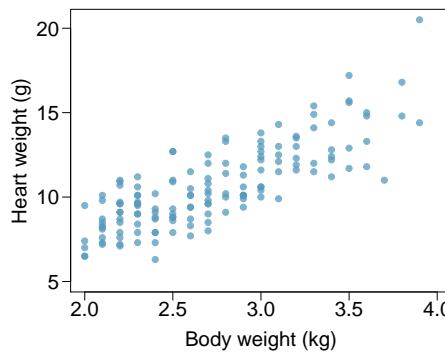
(a) Write out the linear model.

(b) Interpret the intercept.

(c) Interpret the slope.

(d) Interpret R^2.

(e) Calculate the correlation coefficient.
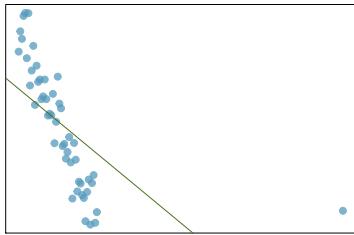
8.26 Cats, Part I. The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$s = 1.452$ $\quad R^2 = 64.66\%$ $\quad R^2_{adj} = 64.41\%$
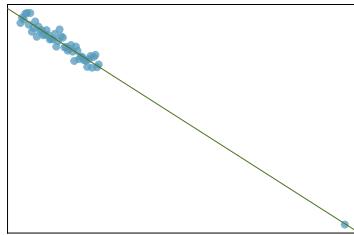
(a) Write out the linear model.

(b) Interpret the intercept.

(c) Interpret the slope.

(d) Interpret R^2.

(e) Calculate the correlation coefficient.
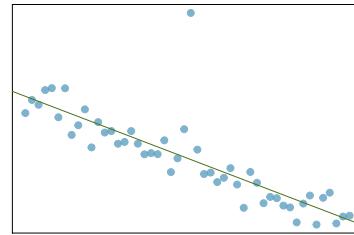
8.27 Outliers, Part I. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.
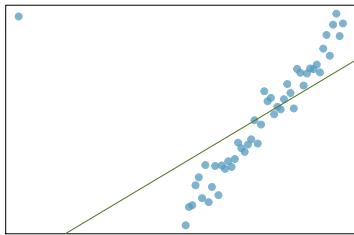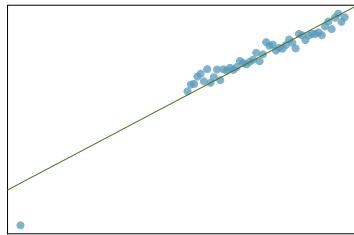
(a)   (b)   (c)
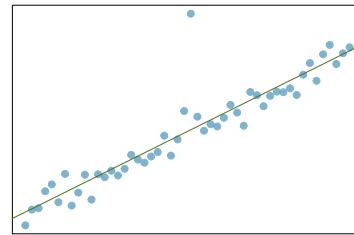
8.28 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.
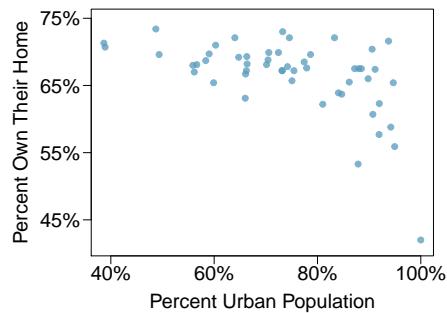
(a)   (b)   (c)

8.29   Urban homeowners, Part I. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas.[21] There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.

(a) Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas.

(b) The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of an outlier is this observation?

8.30   Crawling babies, Part II. Exercise 8.12 introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53°F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an influential point?

---

[21]United States Census Bureau, 2010 Census Urban and Rural Classification and Urban Area Criteria and Housing Characteristics: 2010.

8.3 Inference for the slope of a regression line

Here we encounter our last confidence interval and hypothesis test procedures, this time for making inferences about the slope of the population regression line. We can use this to answer questions such as the following:

- Is the unemployment rate a significant linear predictor for the loss of the President's party in the House of Representatives?

- On average, how much less in college gift aid do students receive when their parents earn an additional $1000 in income?

Learning objectives

1. Recognize that the slope of the sample regression line is a point estimate and has an associated standard error.

2. Be able to read the results of computer regression output and identify the quantities needed for inference for the slope of the regression line, specifically the slope of the sample regression line, the SE of the slope, and the degrees of freedom.

3. State and verify whether or not the conditions are met for inference on the slope of the regression line based using the t-distribution.

4. Carry out a complete confidence interval procedure for the slope of the regression line.

5. Carry out a complete hypothesis test for the slope of the regression line.

6. Distinguish between when to use the t-test for the slope of a regression line and when to use the matched pairs t-test for a mean of differences.

8.3.1 The role of inference for regression parameters

Previously, we found the equation of the regression line for predicting gift aid from family income at Elmhurst College. The slope, b, was equal to -0.0431. This is the slope for our sample data. However, the sample was taken from a larger population. We would like to use the slope computed from our sample data to estimate the slope of the population regression line.

The equation for the population regression line can be written as

$$\mu_y = \alpha + \beta x$$

Here, α and β represent two model parameters, namely the y-intercept and the slope of the true or population regression line. (This use of α and β have nothing to do with the α and β we used previously to represent the probability of a Type I Error and Type II Error!) The parameters α and β are estimated using data. We can look at the equation of the regression line calculated from a particular data set:

$$\hat{y} = a + bx$$

and see that a and b are point estimates for α and β, respectively. If we plug in the values of a and b, the regression equation for predicting gift aid based on family income is:

$$\hat{y} = 24.3193 - 0.0431x$$

The slope of the sample regression line, -0.0431, is our best estimate for the slope of the population regression line, but there is variability in this estimate since it is based on a sample. A different sample would produce a somewhat different estimate of the slope. The standard error of the slope tells us the typical variation in the slope of the sample regression line and the typical error in using this slope to estimate the slope of the population regression line.

We would like to construct a 95% confidence interval for β, the slope of the population regression line. As with means, inference for the slope of a regression line is based on the t-distribution.

INFERENCE FOR THE SLOPE OF A REGRESSION LINE

Inference for the slope of a regression line is based on the t-distribution with $n - 2$ degrees of freedom, where n is the number of paired observations.

Once we verify that conditions for using the t-distribution are met, we will be able to construct the confidence interval for the slope using a critical value $t^{\star}$ based on $n - 2$ degrees of freedom. We will use a table of the regression summary to find the point estimate and standard error for the slope.

8.3.2 Conditions for the least squares line

Conditions for inference in the context of regression can be more complicated than when dealing with means or proportions.

Inference for parameters of a regression line involves the following assumptions:

Linearity. The true relationship between the two variables follows a linear trend. We check whether this is reasonable by examining whether the data follows a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 8.22), an advanced regression method from another book or later course should be applied.

Nearly normal residuals. For each x-value, the residuals should be nearly normal. When this assumption is found to be unreasonable, it is usually because of outliers or concerns about influential points. An example which suggestions non-normal residuals is shown in the second panel of Figure 8.22. If the sample size $n \geq 30$, then this assumption is not necessary.

Constant variability. The variability of points around the true least squares line is constant for all values of x. An example of non-constant variability is shown in the third panel of Figure 8.22.

Independent. The observations are independent of one other. The observations can be considered independent when they are collected from a random sample or randomized experiment. Be careful of data collected sequentially in what is called a **time series**. An example of data collected in such a fashion is shown in the fourth panel of Figure 8.22.

We see in Figure 8.22, that patterns in the residual plots suggest that the assumptions for regression inference are not met in those four examples. In fact, identifying nonlinear trends in the data, outliers, and non-constant variability in the residuals are often easier to detect in a residual plot than in a scatterplot.

We note that the second assumption regarding nearly normal residuals is particularly difficult to assess when the sample size is small. We can make a graph, such as a histogram, of the residuals, but we cannot expect a small data set to be nearly normal. All we can do is to look for excessive skew or outliers. Outliers and influential points in the data can be seen from the residual plot as well as from a histogram of the residuals.
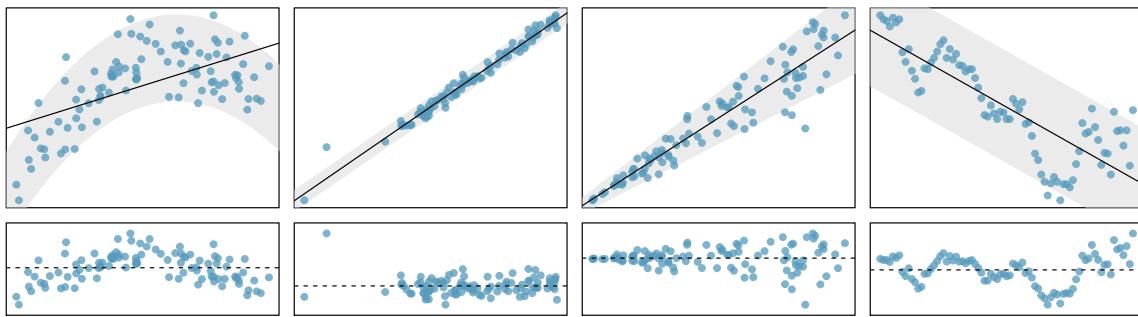
Figure 8.22: Four examples showing when the inference methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not fit the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.

CONDITIONS FOR INFERENCE ON THE SLOPE OF A REGRESSION LINE

1. The data is collected from a random sample or randomized experiment.

2. The residual plot appears as a random cloud of points and does not have any patterns or significant outliers that would suggest that the linearity, nearly normal residuals, constant variability, or independence assumptions are unreasonable.

8.3.3 Constructing a confidence interval for the slope of a regression line

We would like to construct a confidence interval for the slope of the regression line for predicting gift aid based on family income for *all* freshmen at Elmhurst college.

Do conditions seem to be satisfied? We recall that the 50 freshmen in the sample were randomly chosen, so the observations are independent. Next, we need to look carefully at the scatterplot and the residual plot.

ALWAYS CHECK CONDITIONS

Do not blindly apply formulas or rely on regression output; always first look at a scatterplot or a residual plot. If conditions for fitting the regression line are not met, the methods presented here should not be applied.

The scatterplot seems to show a linear trend, which matches the fact that there is no curved trend apparent in the residual plot. Also, the standard deviation of the residuals is mostly constant for different x values and there are no outliers or influential points. There are no patterns in the residual plot that would suggest that a linear model is not appropriate, so the conditions are reasonably met. We are now ready to calculate the 95% confidence interval.
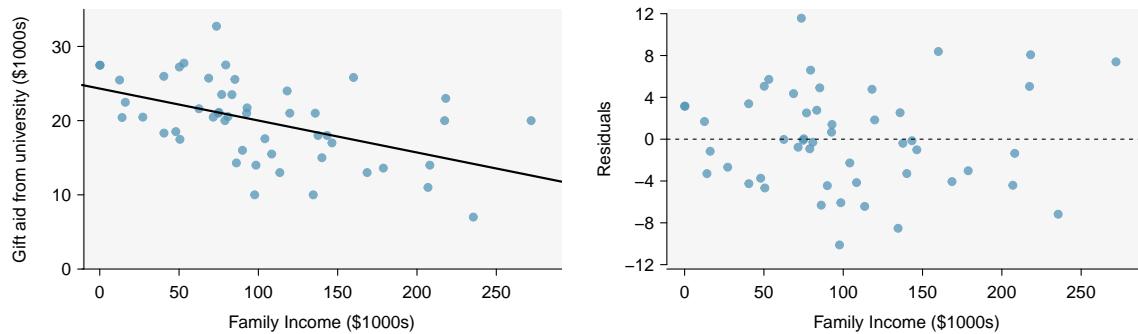
Figure 8.23: Left: Scatterplot of gift aid versus family income for 50 freshmen at Elmhurst college. Right: Residual plot for the model shown in left panel.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.24: Summary of least squares fit for the Elmhurst College data, where we are predicting gift aid by the university based on the family income of students.

EXAMPLE 8.26

Construct a 95% confidence interval for the slope of the regression line for predicting gift aid from family income at Elmhurst college.

——————

As usual, the confidence interval will take the form:

$$\text{point estimate} \pm \text{critical value} \times SE \text{ of estimate}$$

The point estimate for the slope of the population regression line is the slope of the sample regression line: -0.0431. The standard error of the slope can be read from the table as 0.0108. Note that we do not need to divide 0.0108 by the square root of n or do any further calculations on 0.0108; 0.0108 is the SE of the slope. Note that the value of t given in the table refers to the test statistic, not to the critical value $t^\star$. To find $t^\star$ we can use a t-table. Here $n = 50$, so $df = 50 - 2 = 48$. Using a t-table, we round down to row $df = 40$ and we estimate the critical value $t^\star = 2.021$ for a 95% confidence level. The confidence interval is calculated as:

$$-0.0431 \pm 2.021 \times 0.0108$$
$$= (-0.065, -0.021)$$

Note: $t^\star$ using exactly 48 degrees of freedom is equal to 2.01 and gives the same interval of $(-0.065, -0.021)$.

EXAMPLE 8.27

Intepret the confidence interval in context. What can we conclude?

——————

We are 95% confident that the slope of the population regression line, the true average change in gift aid for each additional $1000 in family income, is between $-\$0.065$ thousand dollars and $-\$0.021$ thousand dollars. That is, we are 95% confident that, on average, when family income is $1000 higher, gift aid is between $21 and $65 *lower*.

Because the entire interval is negative, we have evidence that the slope of the population regression line is less than 0. In other words, we have evidence that there is a significant negative linear relationship between gift aid and family income.
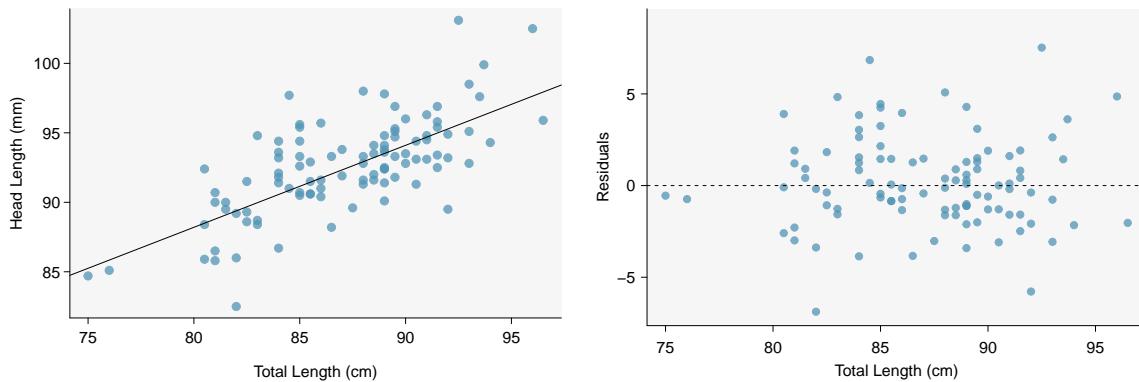
Figure 8.25: Left: Scatterplot of head length versus total length for 104 brushtail possums. Right: Residual plot for the model shown in left panel.

CONSTRUCTING A CONFIDENCE INTERVAL FOR THE SLOPE OF REGRESSION LINE

To carry out a complete confidence interval procedure to estimate the slope of the population regression line β,

Identify: Identify the parameter and the confidence level, C%.

The parameter will be a slope of the population regression line, e.g. the slope of the population regression line relating air quality index to average rainfall per year for each city in the United States.

Choose: Choose the correct interval procedure and identify it by name.

Here we use choose the **t-interval for the slope**.

Check: Check conditions for using a t-interval for the slope.

1. Data come from a random sample or randomized experiment.
2. The residual plot shows no pattern implying that a linear model is reasonable.

    More specifically, the residuals should be independent, nearly normal (or $n \geq 30$),

    and have constant standard deviation.

Calculate: Calculate the confidence interval and record it in interval form.

point estimate $\pm$ $t^{\star} \times SE$ of estimate, $\quad df = n - 2$

> point estimate: the slope b of the sample regression line
>
> SE of estimate: SE of slope (find using computer output)
>
> $t^{\star}$: use a t-distribution with $df = n - 2$ and confidence level C%

(____, ____)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

We are C% confident that the true *slope* of the regression line, the average change in [y] for each unit increase in [x], is between ____ and ____. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.

EXAMPLE 8.28

The regression summary below shows statistical software output from fitting the least squares regression line for predicting head length from total length for 104 brushtail possums. The scatterplot and residual plot are shown above.

```
Predictor          Coef        SE Coef    T        P
Constant          42.70979     5.17281    8.257    5.66e-13
total_length       0.57290     0.05933    9.657    4.68e-16

S = 2.595    R-Sq = 47.76%    R-Sq(adj) = 47.25%
```

Construct a 95% confidence interval for the slope of the regression line. Is there convincing evidence that there is a positive, linear relationship between head length and total length? Use the five step framework to organize your work.

——————

Identify: The parameter of interest is the slope of the population regression line for predicting head length from body length. We want to estimate this at the 95% confidence level.

Choose: Because the parameter to be estimated is the slope of a regression line, we will use the t-interval for the slope.

Check: These data come from a random sample. The residual plot shows no pattern. In general, the residuals have constant standard deviation and there are no outliers or influential points. Also $n = 104 \geq 30$ so some skew in the residuals would be acceptable. A linear model is reasonable here.

Calculate: We will calculate the interval: point estimate $\pm\ t^\star \times SE$ of estimate

We read the slope of the sample regression line and the corresponding SE from the table. The point estimate is $b = 0.57290$. The SE of the slope is 0.05933, which can be found next to the slope of 0.57290. The degrees of freedom is $df = n - 2 = 104 - 2 = 102$. As before, we find the critical value $t^\star$ using a t-table (the $t^\star$ value is not the same as the T-statistic for the hypothesis test). Using the t-table at row $df = 100$ (round down since 102 is not on the table) and confidence level 95%, we get $t^\star = 1.984$.

So the 95% confidence interval is given by:

$$0.57290 \pm 1.984 \times 0.05933$$
$$(0.456, 0.691)$$

Conclude: We are 95% confident that the slope of the population regression line is between 0.456 and 0.691. That is, we are 95% confident that the true average *increase* in head length for each additional cm in total length is between 0.456 mm and 0.691 mm. Because the interval is entirely above 0, we do have evidence of a positive linear association between the head length and body length for brushtail possums.

8.3.4 Calculator: the linear regression t-interval for the slope

We will rely on regression output from statistical software when constructing confidence intervals for the slope of a regression line. We include calculator instructions here simply for completion.

TI-84: T-INTERVAL FOR β

Use STAT, TESTS, LinRegTInt.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose G: LinRegTInt.
   - This test is not built into the TI-83.
4. Let Xlist be L1 and Ylist be L2. (Don't forget to enter the x and y values in L1 and L2 before doing this interval.)
5. Let Freq be 1.
6. Enter the desired confidence level.
7. Leave RegEQ blank.
8. Choose Calculate and hit ENTER, which returns:

(__,__)	the confidence interval
b	b, the slope of best fit line of the sample data
df	degrees of freedom associated with this confidence interval
s	standard deviation of the residuals (not the same as SE of the slope)
a	a, the y-intercept of the best fit line of the sample data
r²	R^2, the explained variance
r	r, the correlation coefficient

8.3.5 Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. Figure 8.26 shows these data and the least-squares regression line:

$$\% \text{ change in House seats for President's party}$$
$$= -7.36 - 0.89 \times (\text{unemployment rate})$$

We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Republicans in 2018) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or the normality of residuals. While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found.
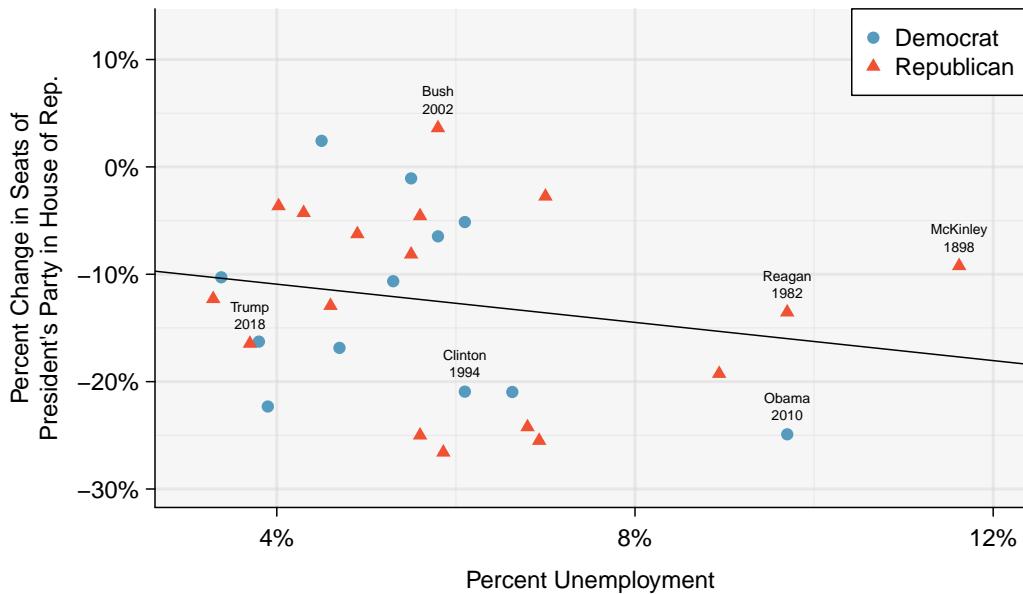
Figure 8.26: The percent change in House seats for the President's party in each election from 1898 to 2018 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data. Explore this data set on Tableau Public.

GUIDED PRACTICE 8.29

The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?[22]

There is a negative slope in the line shown in Figure 8.26. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the "true" linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate? We can frame this investigation as a statistical hypothesis test:

H_0: $\beta = 0$. The true linear model has slope zero.

H_A: $\beta < 0$. The true linear model has a slope less than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives.

We would reject H_0 in favor of H_A if the data provide strong evidence that the slope of the population regression line is less than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value. Before we calculate these quantities, how good are we at visually determining from a scatterplot when a slope is significantly less than or greater than 0? And why do we tend to use a 0.05 significance level as our cutoff? Try out the following activity which will help answer these questions.

TESTING FOR THE SLOPE USING A CUTOFF OF 0.05

What does it mean to say that the slope of the population regression line is significantly greater than 0? And why do we tend to use a cutoff of $\alpha = 0.05$? This 5-minute interactive task will explain:

www.openintro.org/why05

---

[22]We will provide two considerations. Each of these points would have very high leverage on any least-squares regression line, and years with such high unemployment may not help us understand what would happen in other years where the unemployment is only modestly high. On the other hand, these are exceptional cases, and we would be discarding important information if we exclude them from a final analysis.

8.3.6 Understanding regression output from software

The residual plot shown in Figure 8.27 shows no pattern that would indicate that a linear model is inappropriate. Therefore we can carry out a test on the population slope using the sample slope as our point estimate. Just as for other point estimates we have seen before, we can compute a standard error and test statistic for b. The test statistic T follows a t-distribution with $n - 2$ degrees of freedom.
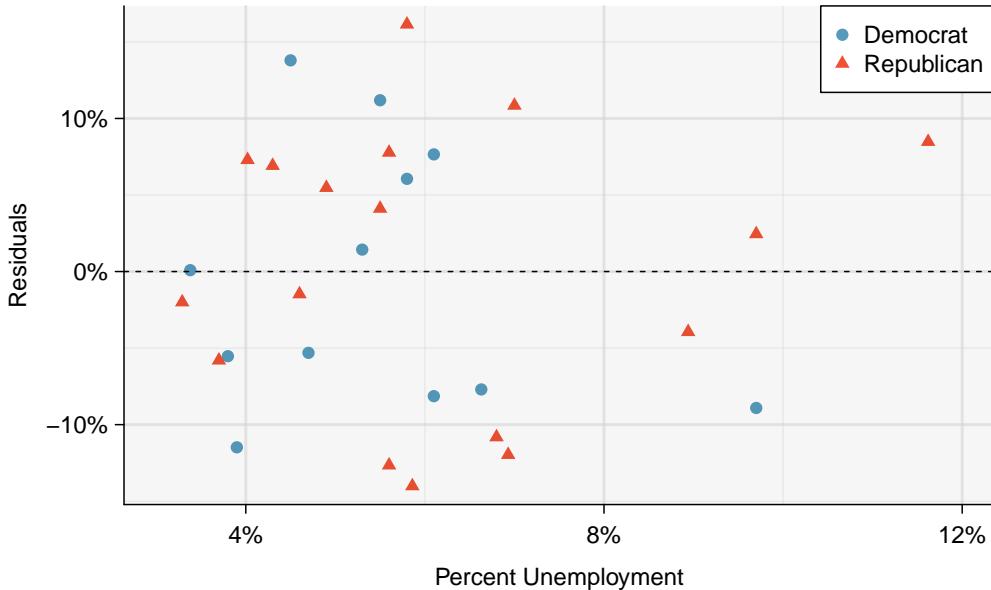

Figure 8.27: The residual plot shows no pattern that would indicate that a linear model is inappropriate. Explore this data set on Tableau Public.

HYPOTHESIS TESTS ON THE SLOPE OF THE REGRESSION LINE

Use a t-test with $n - 2$ degrees of freedom when performing a hypothesis test on the slope of a regression line.

We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Figure 8.28 shows software output for the least squares regression line in Figure 8.26. The row labeled *unemp* represents the information for the slope, which is the coefficient of the unemployment variable.

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	-7.3644	5.1553	-1.43	0.1646
unemp	-0.8897	0.8350	-1.07	0.2961

Figure 8.28: Least squares regression summary for the percent change in seats of president's party in House of Reprepsentatives based on percent unemployment.
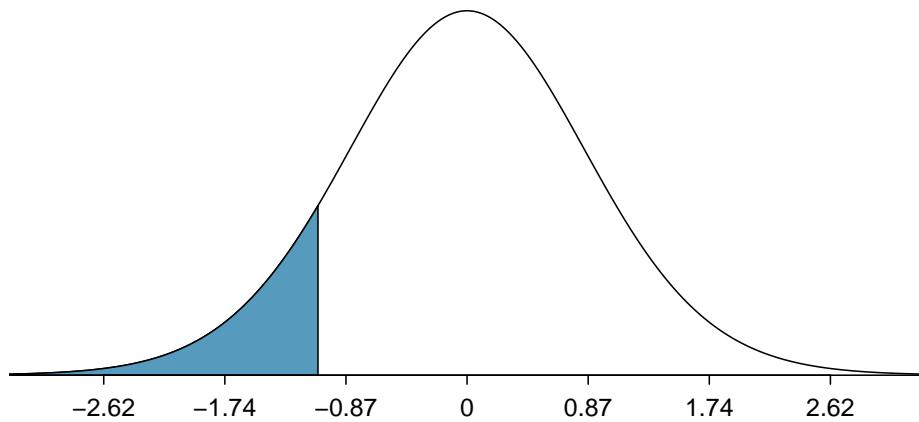
Figure 8.29: The distribution shown here is the sampling distribution for b, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President's party during a midterm election.

EXAMPLE 8.30

What do the first column of numbers in the regression summary represent?

The entries in the first column represent the least squares estimates for the y-intercept and slope, a and b respectively. Using this information, we could write the equation for the least squares regression line as

$$\hat{y} = -7.3644 - 0.8897x$$

where y in this case represents the percent change in the number of seats for the president's party, and x represents the unemployment rate.

We previously used a test statistic T for hypothesis testing in the context of means. Regression is very similar. Here, the point estimate is $b = -0.8897$. The SE of the estimate is 0.8350, which is given in the second column, next to the estimate of b. This SE represents the typical error when using the slope of the sample regression line to estimate the slope of the population regression line.

The null value for the slope is 0, so we now have everything we need to compute the test statistic. We have:

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}} = \frac{-0.8897 - 0}{0.8350} = -1.07$$

This value corresponds to the T-score reported in the regression output in the third column along the *unemp* row.

EXAMPLE 8.31

In this example, the sample size $n = 27$. Identify the degrees of freedom and p-value for the hypothesis test.

The degrees of freedom for this test is $n - 2$, or $df = 27 - 2 = 25$. We could use a table or a calculator to find the probability of a value less than -1.07 under the t-distribution with 25 degrees of freedom. However, the two-side p-value is given in Figure 8.28, next to the corresponding t-statistic. Because we have a one-sided alternate hypothesis, we take half of this. The p-value for the test is $\frac{0.2961}{2} = 0.148$.

Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not provide convincing evidence that a higher unemployment rate is associated with a larger loss for the President's party in the House of Representatives in midterm elections.

DON'T CARELESSLY USE THE P-VALUE FROM REGRESSION OUTPUT

The last column in regression output often lists p-values for one particular hypothesis: a two-sided test where the null value is zero. If your test is one-sided and the point estimate is in the direction of H_A, then you can halve the software's p-value to get the one-tail area. If neither of these scenarios match your hypothesis test, be cautious about using the software output to obtain the p-value.

HYPOTHESIS TEST FOR THE SLOPE OF REGRESSION LINE

To carry out a complete hypothesis test for the claim that there is no linear relationship between two numerical variables, i.e. that $\beta = 0$,

Identify: Identify the hypotheses and the significance level, α.

H_0: $\beta = 0$

H_A: $\beta \neq 0$; H_A: $\beta > 0$;   or H_A: $\beta < 0$

Choose: Choose the correct test procedure and identify it by name.

Here we choose the **t-test for the slope**.

Check: Check conditions for using a t-test for the slope.

1. Data come from a random sample or randomized experiment.
2. The residual plot shows no pattern implying that a linear model is reasonable.
   More specifically, the residuals should be independent, nearly normal (or $n \geq 30$), and have constant standard deviation.

Calculate: Calculate the t-statistic, df, and p-value.

$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}, \quad df = n - 2$

point estimate: the slope b of the sample regression line

SE of estimate: SE of slope (find using computer output)

null value: 0

p-value = (based on the t-statistic, the df, and the direction of H_A)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that $[H_A$ in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that $[H_A$ in context].

EXAMPLE 8.32

The regression summary below shows statistical software output from fitting the least squares regression line for predicting gift aid based on family income for 50 freshman students at Elmhurst College. The scatterplot and residual plot were shown in Figure 8.23.

```
Predictor          Coef       SE Coef   T        P
Constant           24.31933   1.29145   18.831   < 2e-16
family_income      -0.04307   0.01081   -3.985   0.000229


S = 4.783    R-Sq = 24.86%    R-Sq(adj) = 23.29%
```

Do these data provide convincing evidence that there is a negative, linear relationship between family income and gift aid? Carry out a complete hypothesis test at the 0.05 significance level. Use the five step framework to organize your work.

———————

Identify: We will test the following hypotheses at the $\alpha = 0.05$ significance level.
  H_0: $\beta = 0$. There is no linear relationship.
  H_A: $\beta < 0$. There is a negative linear relationship.

Here, β is the slope of the population regression line for predicting gift aid from family income at Elmhurst College.

Choose: Because the hypotheses are about the slope of a regression line, we choose the t-test for a slope.

Check: The data come from a random sample. Also, the residual plot shows that the residuals have constant variance and no outliers or influential points (and $n = 50 \geq 30$). The lack of any pattern in the residual plot indicates that a linear model is appropriate.

Calculate: We will calculate the t-statistic, degrees of freedom, and the p-value.

$$T = \frac{\text{point estimate} \; - \; \text{null value}}{SE \text{ of estimate}}$$

We read the slope of the sample regression line and the corresponding SE from the table.
The point estimate is: $b = -0.04307$.
The SE of the slope is: $SE = 0.01081$.

$$T = \frac{-0.04307 - 0}{0.01081} = -3.985$$

Because H_A uses a less than sign ($<$), meaning that it is a lower-tail test, the p-value is the area to the *left* of $t = -3.985$ under the t-distribution with $50 - 2 = 48$ degrees of freedom. The p-value $= \frac{1}{2}(0.000229) \approx 0.0001$.

Conclude: The p-value of 0.0001 is < 0.05, so we reject H_0; there is sufficient evidence that there is a negative linear relationship between family income and gift aid at Elmhurst College.

8.3.7 Calculator: the t-test for the slope

When performing this type of inference, we generally make use of regression output that provides us with the necessary quantities: b and SE of b. The calculator functions below require knowing all of the data and are, therefore, rarely used. We describe them here for the sake of completion.

TI-83/84: LINEAR REGRESSION T-TEST ON β

Use STAT, TESTS, LinRegTTest.

1. Choose STAT.

2. Right arrow to TESTS.

3. Down arrow and choose F:LinRegTTest. (On TI-83 it is E:LinRegTTest).

4. Let Xlist be L1 and Ylist be L2. (Don't forget to enter the x and y values in L1 and L2 before doing this test.)

5. Let Freq be 1.

6. Choose $\neq$, $<$, or $>$ to correspond to H_A.

7. Leave RegEQ blank.

8. Choose Calculate and hit ENTER, which returns:

t	t statistic	b	b, slope of the line
p	p-value	s	st. dev. of the residuals
df	degrees of freedom for the test	r²	R^2, explained variance
a	a, y-intercept of the line	r	r, correlation coefficient

CASIO FX-9750GII: LINEAR REGRESSION T-TEST ON β

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).

2. Enter your data into 2 lists.

3. Select TEST (F3), t (F2), and REG (F3).

4. If needed, update the sidedness of the test and the XList and YList lists. The Freq should be set to 1.

5. Hit EXE, which returns:

t	t statistic	b	b, slope of the line
p	p-value	s	st. dev. of the residuals
df	degrees of freedom for the test	r	r, correlation coefficient
a	a, y-intercept of the line	r²	R^2, explained variance

EXAMPLE 8.33

Why does the calculator test include the symbol ρ when choosing the direction of the alternate hypothesis?

———

Recall the we used the letter r to represent correlation. The Greek letter $\rho = 0$ represents the correlation for the entire population. The slope $b = r\frac{s_y}{s_x}$. If the slope of the population regression line is zero, the correlation for the population must also be zero. For this reason, the t-test for $\beta = 0$ is equivalent to a test for $\rho = 0$.

8.3.8   Which inference procedure to use for paired data?

In Section 7.2.4, we looked at a set of paired data involving the price of textbooks for UCLA courses at the UCLA Bookstore and on Amazon. The left panel of Figure 8.30 shows the difference in price (UCLA Bookstore − Amazon) for each book. Because we have two data points on each textbook, it also makes sense to construct a scatterplot, as seen in the right panel of Figure 8.30.
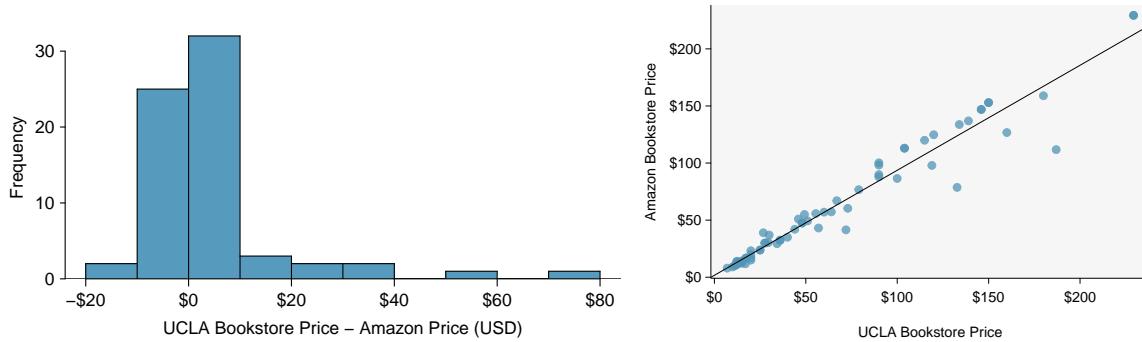

Figure 8.30: Left: histogram of the difference (UCLA Bookstore - Amazon) in price for each book sampled. Right: scatterplot of Amazon Price versus UCLA Bookstore price.

EXAMPLE 8.34

What additional information does the scatterplot provide about the price of textbooks at UCLA Bookstore and on Amazon?

With a scatterplot, we see the *relationship* between the variables. We can see when UCLA Bookstore price is larger, whether Amazon price tends to be larger. We can consider the strength of the correlation and we can plot the linear regression equation.

EXAMPLE 8.35

Which test should we do if we want to check whether:

1. prices for textbooks for UCLA courses are *higher* at the UCLA Bookstore than on Amazon

2. there is a significant, positive linear relationship between UCLA Bookstore price and Amazon price?

In the first case, we are interested in whether the differences (UCLA Bookstore − Amazon) are, on average, greater than 0, so we would do a matched pairs t-test for a mean of differences. In the second case, we are interested in whether the slope is significantly greater than 0, so we would do a t-test for the slope of a regression line.

Likewise, a matched pairs t-interval for a mean of differences would provide an interval of reasonable values for mean of the differences for all UCLA textbooks, whereas a t-interval for the slope would provide an interval of reasonable values for the slope of the regression line for all UCLA textbooks.

INFERENCE FOR PAIRED DATA

A matched pairs t-interval or t-test for a mean of differences only makes sense when we are asking whether, on average, one variable is *greater* than another (think histogram of the differences). A t-interval or t-test for the slope of a regression line makes sense when we are interested in the linear relationship between them (think scatterplot).

EXAMPLE 8.36

Previously, we looked at the relationship betweeen body length and head length for bushtail possums. We also looked at the relationship between gift aid and family income for freshmen at Elmhurst College. Could we do a matched pairs t-test in either of these scenarios?

———

We have to ask ourselves, does it make sense to ask whether, on average, body length is greater than head length? Similarly, does it make sense to ask whether, on average, gift aid is greater than family income? These don't seem to be meaningful research questions; a matched pairs t-test for a mean of differences would not be useful here.

GUIDED PRACTICE 8.37

A teacher gives her class a pretest and a posttest. Does this result in paired data? If so, which hypothesis test should she use?[23]

———

[23]Yes, there are two observations for each individual, so there is paired data. The appropriate test depends upon the question she wants to ask. If she is interested in whether, on average, students do better on the posttest than the pretest, should use a matched pairs t-test for a mean of differences. If she is interested in whether pretest score is a significant linear predictor of posttest score, she should do a t-test for the slope. In this situation, both tests could be useful, but which one should be used is dependent on the teacher's research question.

Section summary
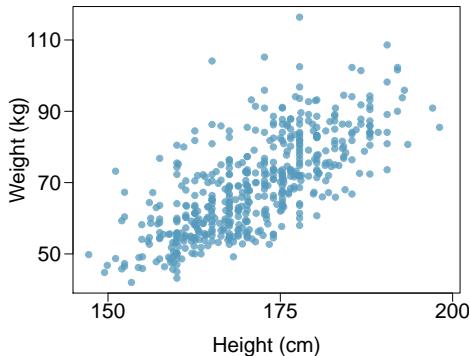
In Chapter 6, we used a χ^2 test of independence to test for association between two categorical variables. In this section, we test for association/correlation between two numerical variables.

- We use the slope b as a *point estimate* for the slope β of the population regression line. The slope of the population regression line is the true increase/decrease in y for each unit increase in x. If the slope of the population regression line is 0, there is no linear relationship between the two variables.

- Under certain assumptions, the sampling distribution of b is *normal* and the distribution of the standardized test statistic using the standard error of the slope follows a **t-distribution** with $n - 2$ degrees of freedom.

- When there is (x, y) data and the parameter of interest is the slope of the population regression line, e.g. the slope of the population regression line relating air quality index to average rainfall per year for each city in the United States:

    - Estimate β at the C% confidence level using a **t-interval for the slope**.
    - Test H_0: $\beta = 0$ at the α significance level using a **t-test for the slope**.

- The conditions for the t-interval and t-test for the slope of a regression line are the same.

    1. Data come from a random sample or randomized experiment.
    2. The residual plot shows no pattern implying that a linear model is reasonable.
       More specifically, the residuals should be independent, nearly normal (or $n \geq 30$), and have constant standard deviation.

- The confidence interval and test statistic are calculated as follows:

    Confidence interval: point estimate $\pm$ $t^\star \times SE$ of estimate, or

    Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$ and p-value

         point estimate: the slope b of the sample regression line

         SE of estimate: SE of slope (find using computer output)

         $df = n - 2$

- If the confidence interval for the slope of the population regression line estimates the true average increase in the y-variable for each unit increase in the x-variable.

- The t-test for the slope and the matched pairs t-test for a mean of differences both involve *paired*, numerical data. However, the t-test for the slope asks if the two variables have a linear *relationship*, specifically if the *slope* of the population regression line is different from 0. The matched pairs t-test for a mean of differences, on the other hand, asks if the two variables are in some way the *same*, specifically if the *mean* of the population differences is 0.

Exercises
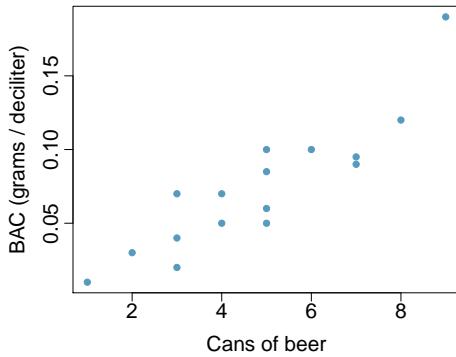
8.31 Body measurements, Part IV. The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-105.0113	7.5394	-13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

(a) Describe the relationship between height and weight.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

8.32 Beer and blood alcohol content. 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 different genders, 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.[24] The scatterplot and regression table summarize the findings.
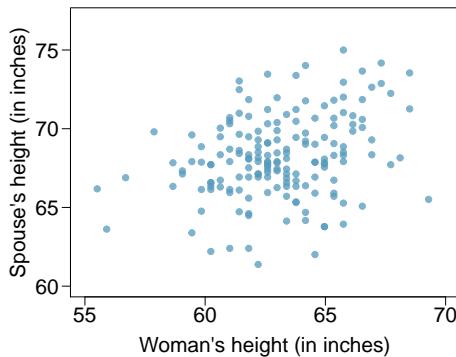

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

(a) Describe the relationship between the number of cans of beer and BAC.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.

(e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

---

[24]J. Malkevitch and L.M. Lesser. *For All Practical Purposes: Mathematical Literacy in Today's World.* WH Freeman & Co, 2008.

8.33 Spouses, Part II. The scatterplot below summarizes womens' heights and their spouses' heights for a random sample of 170 married women in Britain, where both partners' ages are below 65 years. Summary output of the least squares fit for predicting spouse's height from the woman's height is also provided in the table.
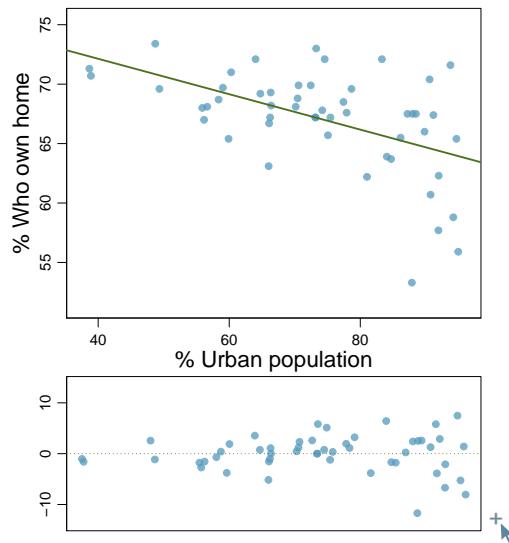

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	43.5755	4.6842	9.30	0.0000
height_spouse	0.2863	0.0686	4.17	0.0000

(a) Is there strong evidence in this sample that taller women have taller spouses? State the hypotheses and include any information used to conduct the test.

(b) Write the equation of the regression line for predicting the height of a woman's spouse based on the woman's height.

(c) Interpret the slope and intercept in the context of the application.

(d) Given that $R^2 = 0.09$, what is the correlation of heights in this data set?

(e) You meet a married woman from Britain who is 5'9" (69 inches). What would you predict her spouse's height to be? How reliable is this prediction?

(f) You meet another married woman from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict her spouse's height? Why or why not?

8.34 Urban homeowners, Part II. Exercise 8.29 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.

(a) For these data, $R^2 = 0.28$. What is the correlation? How can you tell if it is positive or negative?

(b) Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data?

8.35 Murders and poverty, Part II. Exercise 8.25 presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$$s = 5.512 \qquad R^2 = 70.52\% \qquad R^2_{adj} = 68.89\%$$

(a) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

8.36 Babies. Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty-five low birth-weight babies were studied at a Harvard teaching hospital; the investigators calculated the regression of head circumference (measured in centimeters) against gestational age (measured in weeks). The estimated regression line is

$$head\ \widehat{circumfer}ence = 3.91 + 0.78 \times gestational\ age$$

(a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?

(b) The standard error for the coefficient of gestational age is 0. 35, which is associated with $df = 23$. Does the model provide strong evidence that gestational age is significantly associated with head circumference?

8.4   Transformations for skewed data

County population size among the counties in the US is very strongly right skewed. Can we apply a transformation to make the distribution more symmetric? How would such a transformation affect the scatterplot and residual plot when another variable is graphed against this variable? In this section, we will see the power of transformations for very skewed data.

Learning objectives

1. See how a log transformation can bring symmetry to an extremely skewed variable.

2. Recognize that data can often be transformed to produce a linear relationship, and that this transformation often involves log of the y-values and sometimes log of the x-values.

3. Use residual plots to assess whether a linear model for transformed data is reasonable.

8.4.1   Introduction to transformations

EXAMPLE 8.38

Consider the histogram of county populations shown in Figure 8.31(a), which shows extreme skew. What isn't useful about this plot?

Nearly all of the data fall into the left-most bin, and the extreme skew obscures many of the potentially interesting details in the data.

There are some standard transformations that may be useful for strongly right skewed data where much of the data is positive but clustered near zero. A **transformation** is a rescaling of the data using a function. For instance, a plot of the logarithm (base 10) of county populations results in the new histogram in Figure 8.31(b). This data is symmetric, and any potential outliers appear much less extreme than in the original data set. By reigning in the outliers and extreme skew, transformations like this often make it easier to build statistical models against the data.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the population change from 2010 to 2017 against the population in 2010 is shown in Figure 8.32(a). In this first scatterplot, it's hard to decipher any interesting patterns because the population variable is so strongly skewed. However, if we apply a $\log_{10}$ transformation to the population variable, as shown in Figure 8.32(b), a positive association between the variables is revealed. While fitting a line to predict population change (2010 to 2017) from population (in 2010) does not seem reasonable, fitting a line to predict population from $\log_{10}(\text{population})$ does seem reasonable.

Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are commonly used by data scientists. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.
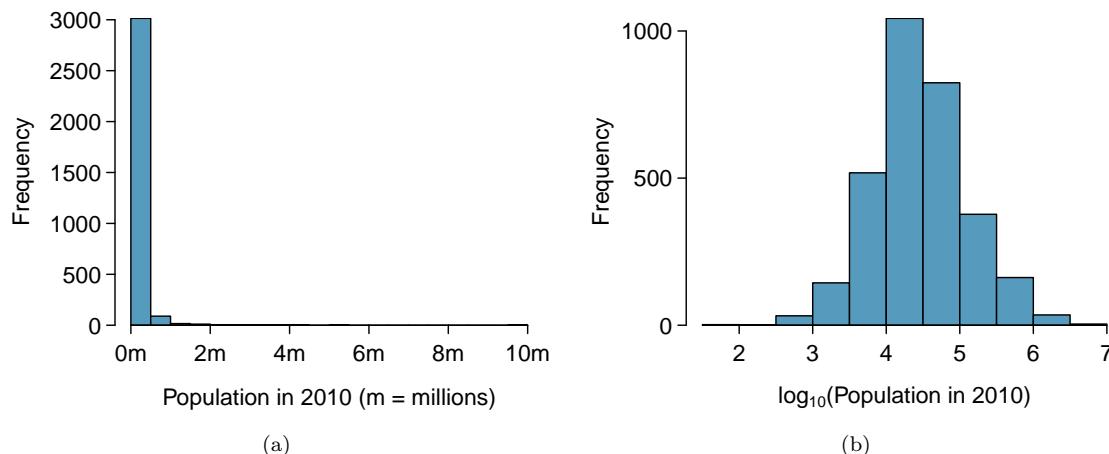
Figure 8.31: (a) A histogram of the populations of all US counties. (b) A histogram of $\log_{10}$-transformed county populations. For this plot, the x-value corresponds to the power of 10, e.g. "4" on the x-axis corresponds to $10^4 = 10{,}000$.
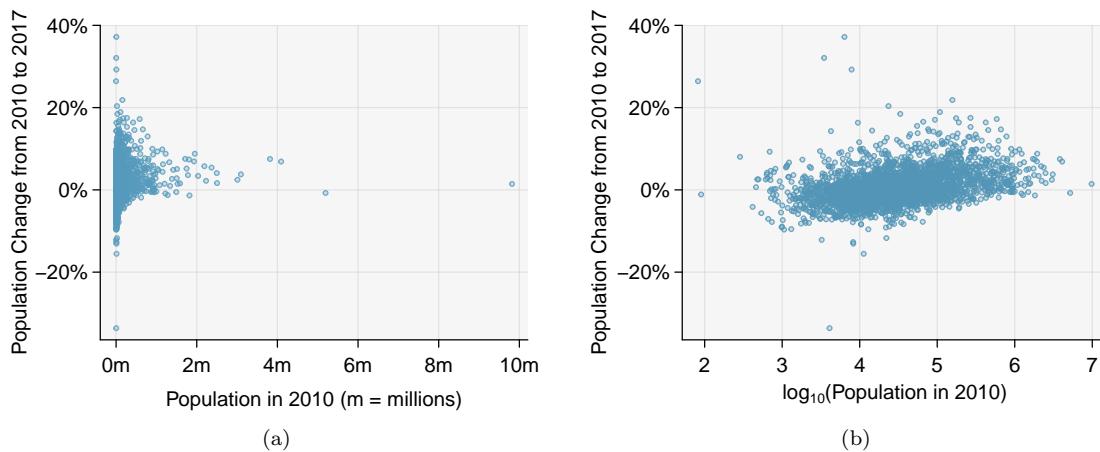

Figure 8.32: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

8.4.2   Transformations to achieve linearity
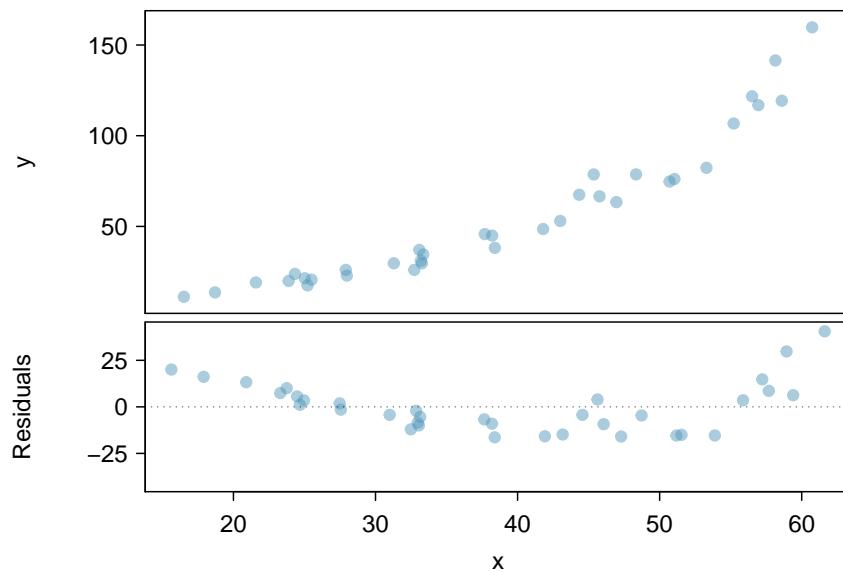
Figure 8.33: Variable y is plotted against x. A nonlinear relationship is evident by the ∪-pattern shown in the residual plot. The curvature is also visible in the original plot.

EXAMPLE 8.39

Consider the scatterplot and residual plot in Figure 8.33. The regression output is also provided. Is the linear model $\hat{y} = -52.3564 + 2.7842x$ a good model for the data?

```
The regression equation is

y = -52.3564 + 2.7842 x

Predictor        Coef    SE Coef           T           P
Constant     -52.3564     7.2757      -7.196        3e-08
x              2.7842     0.1768      15.752      < 2e-16


S = 13.76    R-Sq = 88.26%    R-Sq(adj) = 87.91%
```

We can note the R^2 value is fairly large. However, this alone does not mean that the model is good. Another model might be much better. When assessing the appropriateness of a linear model, we should look at the residual plot. The ∪-pattern in the residual plot tells us the original data is curved. If we inspect the two plots, we can see that for small and large values of x we systematically underestimate y, whereas for middle values of x, we systematically overestimate y. The curved trend can also be seen in the original scatterplot. Because of this, the linear model is not appropriate, and it would not be appropriate to perform a t-test for the slope because the conditions for inference are not met. However, we might be able to use a transformation to linearize the data.

Regression analysis is easier to perform on linear data. When data are nonlinear, we sometimes **transform** the data in a way that makes the resulting relationship linear. The most common **transformation** is log of the y values. Sometimes we also apply a transformation to the x values. We generally use the residuals as a way to evaluate whether the transformed data are more linear. If so, we can say that a better model has been found.

EXAMPLE 8.40

Using the regression output for the transformed data, write the new linear regression equation.

```
The regression equation is

log(y) = 1.722540 + 0.052985 x

Predictor          Coef      SE Coef          T           P
Constant       1.722540     0.056731      30.36     < 2e-16
x              0.052985     0.001378      38.45     < 2e-16

S = 0.1073     R-Sq = 97.82%    R-Sq(adj) = 97.75%
```

The linear regression equation can be written as: $\widehat{\log(y)} = 1.723 + 0.053x$
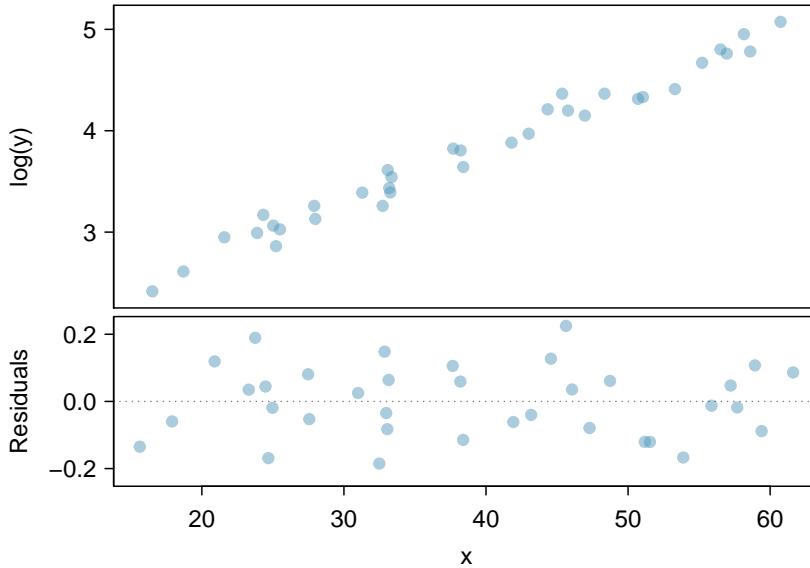
Figure 8.34: A plot of $\log(y)$ against x. The residuals don't show any evident patterns, which suggests the transformed data is well-fit by a linear model.

GUIDED PRACTICE 8.41

Which of the following statements are true? There may be more than one.[25]

(a) There is an apparent linear relationship between x and y.

(b) There is an apparent linear relationship between x and $\widehat{\log(y)}$.

(c) The model provided by Regression I ($\hat{y} = -52.3564 + 2.7842x$) yields a better fit.

(d) The model provided by Regression II ($\widehat{\log(y)} = 1.723 + 0.053x$) yields a better fit.

---
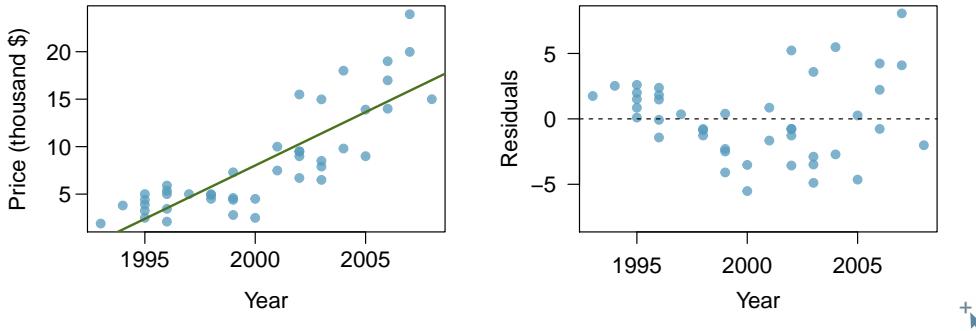
[25]Part (a) is *false* since there is a nonlinear (curved) trend in the data. Part (b) is true. Since the transformed data shows a stronger linear trend, it is a better fit, i.e. Part (c) is *false*, and Part (d) is true.

Section summary
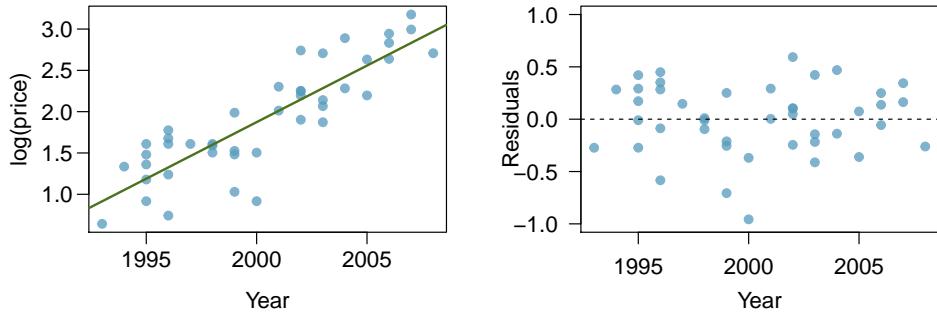
- A **transformation** is a rescaling of the data using a function. When data are very skewed, a log transformation often results in more symmetric data.

- Regression analysis is easier to perform on linear data. When data are nonlinear, we sometimes **transform** the data in a way that results in a linear relationship. The most common transformation is log of the y-values. Sometimes we also apply a transformation to the x-values.

- To assess the model, we look at the **residual plot** of the *transformed* data. If the residual plot of the original data has a pattern, but the residual plot of the transformed data has no pattern, a linear model for the transformed data is reasonable, and the transformed model provides a better fit than the simple linear model.

Exercises

8.37 Used trucks. The scatterplot below shows the relationship between year and price (in thousands of $) of a random sample of 42 pickup trucks. Also shown is a residuals plot for the linear model for predicting price from year.
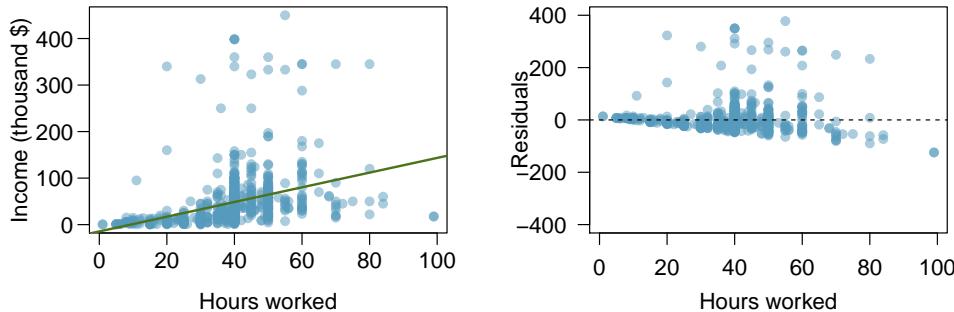
(a) Describe the relationship between these two variables and comment on whether a linear model is appropriate for modeling the relationship between year and price.

(b) The scatterplot below shows the relationship between logged (natural log) price and year of these trucks, as well as the residuals plot for modeling these data. Comment on which model (linear model from earlier or logged model presented here) is a better fit for these data.
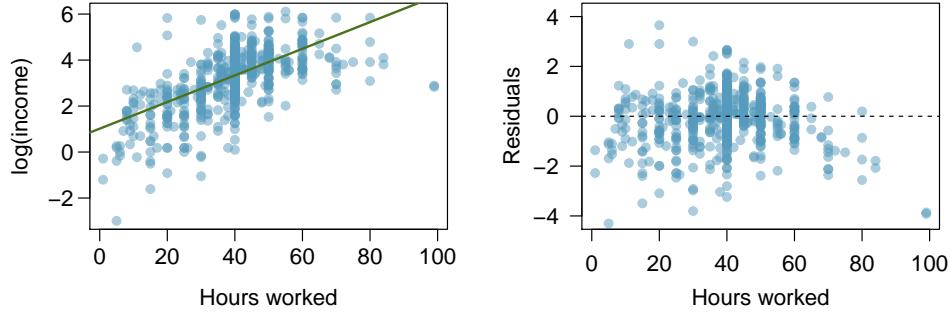
(c) The output for the logged model is given below. Interpret the slope in context of the data.

|	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-271.981	25.042	-10.861	0.000
Year	0.137	0.013	10.937	0.000

8.38   Income and hours worked. The scatterplot below shows the relationship between income and years worked for a random sample of 787 Americans. Also shown is a residuals plot for the linear model for predicting income from hours worked. The data come from the 2012 American Community Survey.[26]

(a) Describe the relationship between these two variables and comment on whether a linear model is appropriate for modeling the relationship between year and price.

(b) The scatterplot below shows the relationship between logged (natural log) income and hours worked, as well as the residuals plot for modeling these data. Comment on which model (linear model from earlier or logged model presented here) is a better fit for these data.

(c) The output for the logged model is given below. Interpret the slope in context of the data.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.017	0.113	9.000	0.000
hrs_work	0.058	0.003	21.086	0.000

---

[26]United States Census Bureau. Summary File. 2012 American Community Survey. U.S. Census Bureau's American Community Survey Office, 2013. Web.

Chapter highlights

This chapter focused on describing the linear association between two numerical variables and fitting a linear model.

- The **correlation coefficient**, r, measures the strength and direction of the linear association between two variables. However, r alone cannot tell us whether data follow a linear trend or whether a linear model is appropriate.

- The **explained variance**, R^2, measures the proportion of variation in the y values explained by a given model. Like r, R^2 alone cannot tell us whether data follow a linear trend or whether a linear model is appropriate.

- Every analysis should begin with *graphing* the data using a **scatterplot** in order to see the association and any deviations from the trend (outliers or influential values). A **residual plot** helps us better see patterns in the data.

- When the data show a linear trend, we fit a **least squares regression line** of the form: $\hat{y} = a + bx$, where a is the y-intercept and b is the slope. It is important to be able to *calculate* a and b using the summary statistics and to *interpret* them in the context of the data.

- A **residual**, $y - \hat{y}$, measures the error for an *individual point*. The **standard deviation of the residuals**, s, measures the typical size of the residuals.

- $\hat{y} = a + bx$ provides the best fit line for the *observed data*. To estimate or hypothesize about the slope of the population regression line, first confirm that the residual plot has no pattern and that a linear model is reasonable, then use a ***t*-interval for the slope** or a ***t*-test for the slope** with $n - 2$ degrees of freedom.

In this chapter we focused on simple linear models with one explanatory variable. More complex methods of prediction, such as multiple regression (more than one explanatory variable) and non-linear regression can be studied in a future course.

Chapter exercises

8.39   True / False. Determine if the following statements are true or false. If false, explain why.

(a) A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

(b) Correlation is a measure of the association between any two variables.
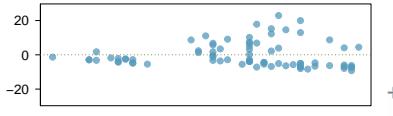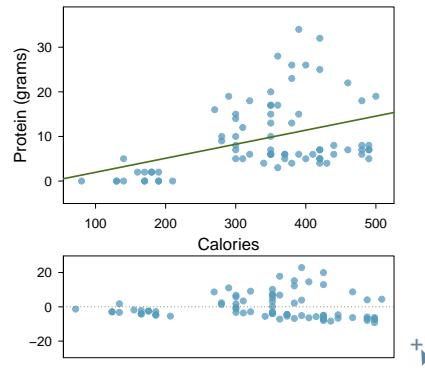
8.40   Cats, Part II. Exercise 8.26 presents regression output from a model for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cat. The model output is also provided below.

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$$s = 1.452 \qquad R^2 = 64.66\% \qquad R^2_{adj} = 64.41\%$$
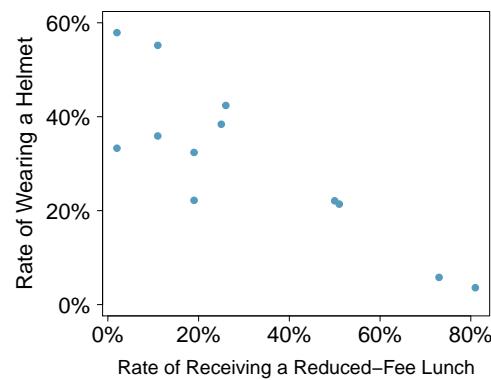
(a) We see that the point estimate for the slope is positive. What are the hypotheses for evaluating whether body weight is positively associated with heart weight in cats?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of body weight, and interpret it in context of the data.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

8.41   Nutrition at Starbucks, Part II. Exercise 8.22 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.
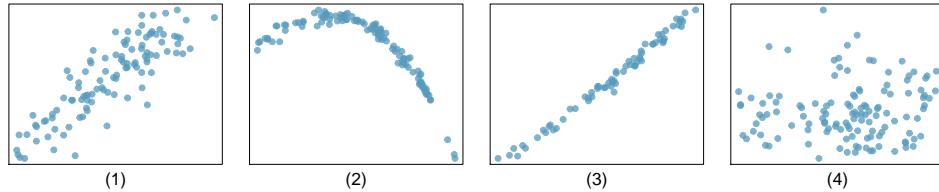

8.42   Helmets and lunches. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (`lunch`) and the percentage of bike riders in the neighborhood wearing helmets (`helmet`). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

(a) If the R^2 for the least-squares regression line for these data is 72%, what is the correlation between `lunch` and `helmet`?

(b) Calculate the slope and intercept for the least-squares regression line for these data.

(c) Interpret the intercept of the least-squares regression line in the context of the application.

(d) Interpret the slope of the least-squares regression line in the context of the application.

(e) What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual in the context of the application.

**8.43 Match the correlation, Part III.** Match each correlation to the corresponding scatterplot.
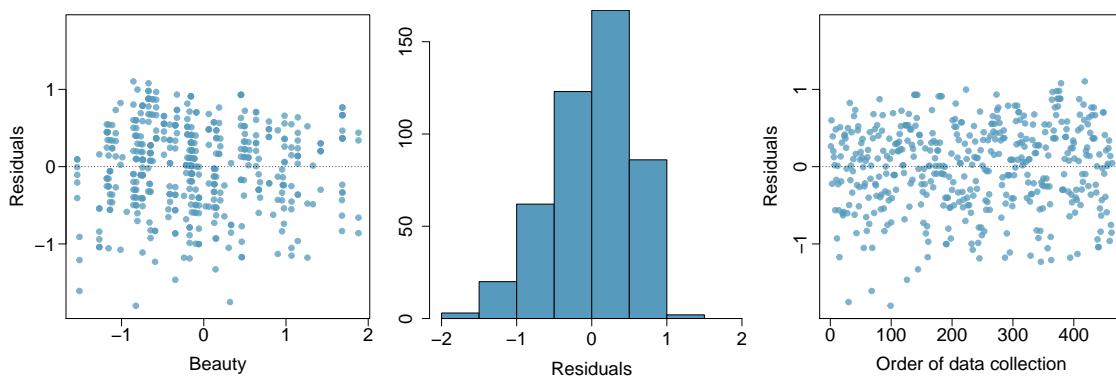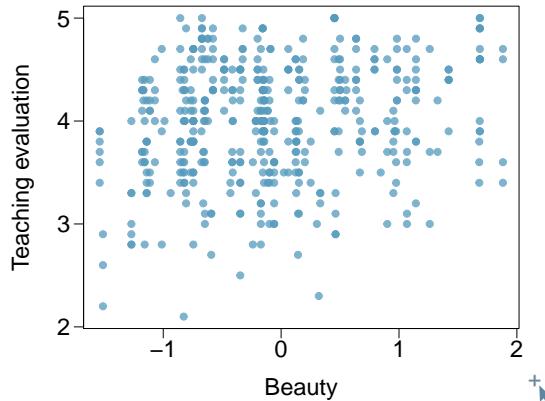
(a) $r = -0.72$

(b) $r = 0.07$

(c) $r = 0.86$

(d) $r = 0.99$

8.44 Rate my professor. Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors.[27] The scatterplot below shows the relationship between these variables, and regression output is provided for predicting teaching evaluation score from beauty score.
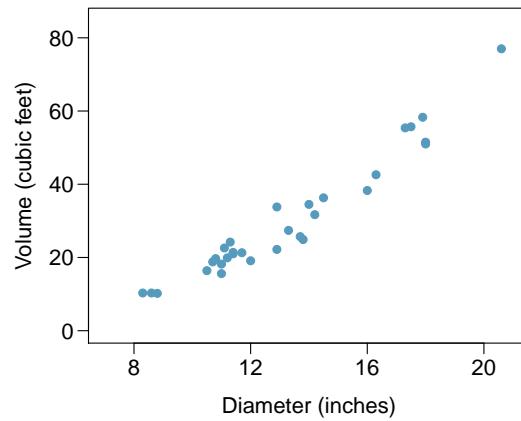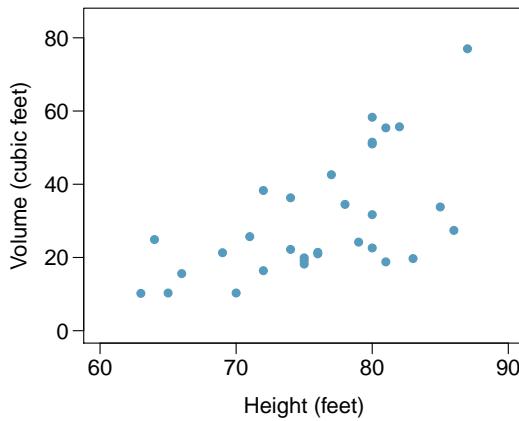
|	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.


[27] Daniel S Hamermesh and Amy Parker. "Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity". In: *Economics of Education Review* 24.4 (2005), pp. 369–376.

8.45 **Trees.** The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.[28]

(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

---

[28]Source: R Dataset, stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html.

Appendix A

Exercise solutions

1 Data collection

1.1 (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. (b) Control: $2/46 = 0.04 \rightarrow 4\%$. (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance.

1.3 (a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM$_{10}$) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables.

1.5 (a) "Does explicitly telling children not to cheat affect their likelihood to cheat?". (b) 160 children between the ages of 5 and 15. (c) Four variables: (1) age (numerical, continuous), (2) sex (categorical), (3) whether they were an only child or not (categorical), (4) whether they cheated or not (categorical).

1.7 Explanatory: acupuncture or not. Response: if the patient was pain free or not.

1.9 (a) $50 \times 3 = 150$. (b) Four continuous numerical variables: sepal length, sepal width, petal length, and petal width. (c) One categorical variable, species, with three levels: *setosa*, *versicolor*, and *virginica*.

1.11 (a) Airport ownership status (public/private), airport usage status (public/private), latitude, and longitude. (b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.

1.13 (a) Population: all births, sample: 143,196 births between 1989 and 1993 in Southern California. (b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

1.15 (a) Population: all asthma patients aged 18-69 who rely on medication for asthma treatment. Sample: 600 such patients. (b) If the patients in this sample, who are likely not randomly sampled, can be considered to be representative of all asthma patients aged 18-69 who rely on medication for asthma treatment, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.17 (a) Observation. (b) Variable. (c) Sample statistic (mean). (d) Population parameter (mean).

1.19 (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.

1.21 (a) Positive, non-linear, somewhat strong. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies, however rise in life expectancy trails off before around 80 years old. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

1.23 (a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method! (b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable. (c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)

1.25 (a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.

1.27 (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population. (b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends. (d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.

1.29 (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Sex: man, woman.

1.31 (a) Exam performance. (b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise). (c) Since the researchers want to ensure equal gender representation, sex will be a blocking variable.

1.33 Need randomization and blinding. One possi-ble outline: (1) Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!) (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double- blind experiment. (Answers may vary.)

1.35 (a) Observational study. (b) Dog: Lucy. Cat: Luna. (c) Oliver and Lily. (d) Positive, as the popularity of a name for dogs increases, so does the popularity of that name for cats.

1.37 (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes, gender. (d) Yes, single blind since the patients were blinded to the treatment they received. (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.
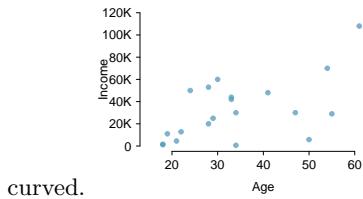
1.39 (a) Non-responders may have a different response to this question, e.g. parents who returned the surveys likely don't have difficulty spending time with their children. (b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might be in a lower socio- economic status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g. these people may go running because they are generally healthier and/or do other exercises.

1.41 (a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, with 3 levels). Response variable: Psychological well-being. (c) No, because the participants were volunteers. (d) Yes, because it was an experiment. (e) The statement should say "evidence" instead of "proof".

1.43 (a) County, state, driver's race, whether the car was searched or not, and whether the driver was arrested or not. (b) All categorical, non-ordinal. (c) Response: whether the car was searched or not. Explanatory: race of the driver.
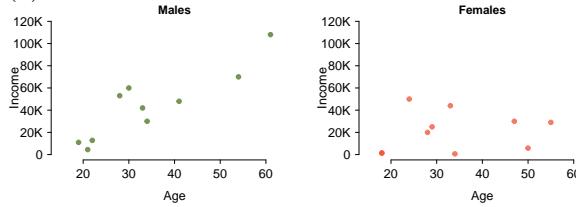
2 Summarizing data

2.1 (a) There is a weak and positive relationship between age and income. With so few points it is difficult to tell the form of the relationship (linear or not) however the relationship does look somewhat

curved.

(b)

(c) For males as age increases so does income, however this pattern is not apparent for females.

2.3 (a)
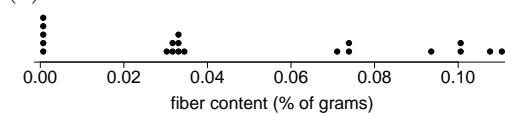
```
0 | 000003333333
0 | 7779
1 | 0011
Legend:  1 | 0 = 10%
```
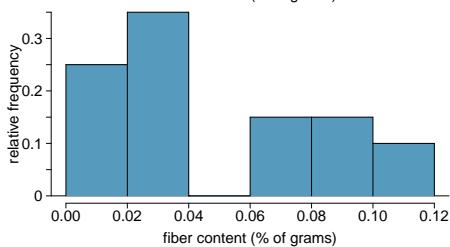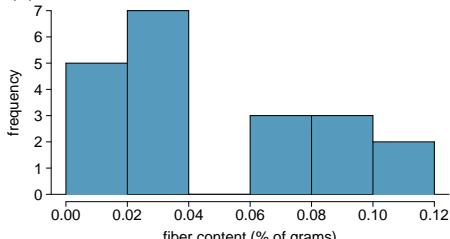
(b)

(c)

(d) 40%

2.5 (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) As-

sociation would still be positive. (c) No, they are not independent. See part (a).

2.7 Both distributions are right skewed and bimodal with modes at 10 and 20 cigarettes; note that people may be rounding their answers to half a pack or a whole pack. The median of each distribution is between 10 and 15 cigarettes. The middle 50% of the data (the IQR) appears to be spread equally in each group and have a width of about 10 to 15. There are potential outliers above 40 cigarettes per day. It appears that respondents who smoke only a few cigarettes (0 to 5) smoke more on the weekdays than on weekends.

2.9 (a) $\bar{x}_{amtWeekends} = 20$, $\bar{x}_{amtWeekdays} = 16$. (b) $s_{amtWeekends} = 0$, $s_{amtWeekdays} = 4.18$. In this very small sample, higher on weekdays.

2.11 Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

2.13 (a) Dist 2 has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13. (b) Dist 1 has a higher mean since $-20 > -40$, and Dist 2 has a higher standard deviation since -40 is farther away from the rest of the data than -20. (c) Dist 2 has a higher mean since all values in this distribution are higher than those in Dist 1, but both distribution have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean since they're both centered at 300, but Dist 2 has a higher standard deviation since the observations are farther from the mean than in Dist 1.

2.15 (a) About 30. (b) Since the distribution is right skewed the mean is higher than the median. (c) Q1: between 15 and 20, Q3: between 35 and 40, IQR: about 20. (d) Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles. Upper fence: $Q3 + 1.5 \times IQR = 37.5 + 1.5 \times 20 = 67.5$; Lower fence: Q1 - 1.5 × IQR $= 17.5 + 1.5 \times 20 = -12.5$; The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

2.17 The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

2.19 (a) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR. (b) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR. (c) The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

2.21 (a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers. (b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

2.23 (a) The distribution is unimodal and symmetric with a mean of about 25 minutes and a standard deviation of about 5 minutes. There does not appear to be any counties with unusually high or low mean travel times. Since the distribution is already unimodal and symmetric, a log transformation is not necessary. (b) Answers will vary. There are pockets of longer travel time around DC, Southeastern NY, Chicago, Minneapolis, Los Angeles, and many other big cities. There is also a large section of shorter average commute times that overlap with farmland in the Midwest. Many farmers' homes are adjacent to their farmland, so their commute would be brief, which may explain why the average commute time for these counties is relatively low.

2.25 (a) We see the order of the categories and the relative frequencies in the bar plot. (b) There are no features that are apparent in the pie chart but not in the bar plot. (c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

2.27 The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.

2.29 (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True. (iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True.
(b) Proportion of all patients who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$
(c) The expected number of heart attacks in the rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7,979}{227,571} \approx 2370$.
(d) (i) H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance. H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems. (ii) A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that rosiglitazone increases the risk of such problems. (iii) In the actual study, we observed 2,593 cardiovascular events in the rosiglitazone group. In the 1,000 simulations under the independence model, we observed somewhat less than 2,593 in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

2.31 (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24+1) = 73.6$. (c) The new score is more than 1 standard deviation away from the previous mean, so increase.

2.33 No, we would expect this distribution to be right skewed. There are two reasons for this: (1) there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV), (2) the standard deviation of the distribution is very large compared to the mean.

2.35 The distribution of ages of best actress winners are right skewed with a median around 30 years. The distribution of ages of best actress winners is also right skewed, though less so, with a median around 40 years. The difference between the peaks of these distributions suggest that best actress winners are typically younger than best actor winners. The ages of best actress winners are more variable than the ages of best actor winners. There are potential outliers on the higher end of both of the distributions.

2.37

3 Probability
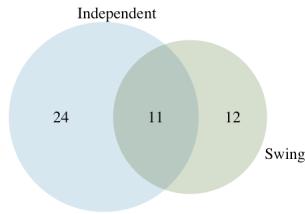
3.1 (a) False. These are independent trials. (b) False. There are red face cards. (c) True. A card cannot be both a face card and an ace.

3.3 (a) 10 tosses. Fewer tosses mean more variability in the sample fraction of heads, meaning there's a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

3.5 (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

3.7 (a) No, there are voters who are both independent and swing voters.
(b)

(c) Each Independent voter is either a swing voter or not. Since 35% of voters are Independents and 11% are both Independent and swing voters, the other 24% must not be swing voters. (d) 0.47. (e) 0.53. (f) P(Independent) × P(swing) = 0.35 × 0.23 = 0.08, which does not equal P(Independent and swing) = 0.11, so the events are dependent.

3.9 (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c). (b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related. (c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

3.11 (a) $0.16 + 0.09 = 0.25$. (b) $0.17 + 0.09 = 0.26$. (c) Assuming that the education level of the husband and wife are independent: $0.25 \times 0.26 = 0.065$. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

3.13 (a) No, but we could if A and B are independent. (b-i) 0.21. (b-ii) 0.79. (b-iii) 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed under independence from part (a). (d) 0.143.

3.15 (a) No, 0.18 of respondents fall into this combination. (b) $0.60 + 0.20 - 0.18 = 0.62$. (c) $0.18/0.20 = 0.9$. (d) $0.11/0.33 \approx 0.33$. (e) No, otherwise the answers to (c) and (d) would be the same. (f) $0.06/0.34 \approx 0.18$.
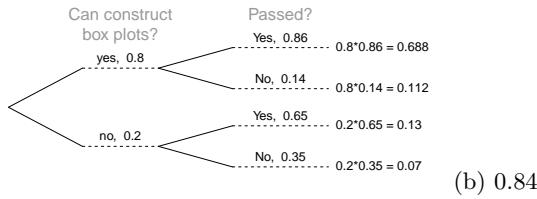
3.17 (a) No. There are 6 females who like Five Guys Burgers. (b) $162/248 = 0.65$. (c) $181/252 = 0.72$. (d) Under the assumption of a dating choices being independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (e) $(252 + 6 - 1)/500 = 0.514$.

3.19 (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sampled from is identical in each draw.

3.21 (a) 2/9. (b) $3/9 = 1/3$. (c) $(3/10) \times (2/9) \approx 0.067$. (d) No. In this small population of marbles, removing one marble meaningfully changes the probability of what might be drawn next.
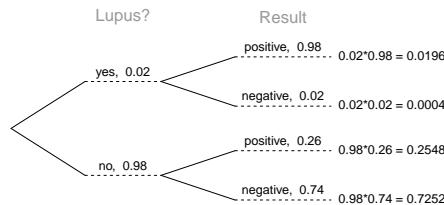
3.23 For 1 leggings (L) and 2 jeans (J), there are three possible orderings: LJJ, JLJ, and JJL. The probability for LJJ is $(5/24) \times (7/23) \times (6/22) = 0.0173$. The other two orderings have the same probability, and these three possible orderings are disjoint events. Final answer: 0.0519.

3.25 (a)

(b) 0.84

3.27 0.0714. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.

3.29 (a) $\binom{5}{1} = 5$. (b) $\binom{5}{4} = 5$. (c) $\binom{5}{3} = 10$. (d) $\binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 10 + 5 + 1 = 16$.

3.31 (a) Yes. The conditions are satisfied: independence, fixed number of trials, either success or failure for each trial, and probability of success being constant across trials. (b) 0.200. (c) 0.200. (d) $0.0024 + 0.0284 + 0.1323 = 0.1631$. (e) $1 - 0.0024 = 0.9976$.

3.33 (a) P(pass) = 0.5, but it should be 0.16. (b) P(pass) = 0.2, instead of 0.16. (c) P(pass) = 0.17, instead of 0.16.

3.35 (a) Starting at row 3 of the random number table, we will read across the table two digits at a time. If the random number is between 00-15, the car will fail the pollution test. If the number is between 16-99, the car will pass the test. (Answers may vary.)

(b) Fleet 1: 18-52-97-32-85-95-29 → P-P-P-P-P-P-P → fleet passes
Fleet 2: 14-96-06-67-17-49-59 → F-P-F-P-P-P-P → fleet fails
Fleet 3: 05-33-67-97-58-11-81 → F-P-P-P-P-F-P → fleet fails
Fleet 4: 23-81-83-21-71-08-50 → P-P-P-P-P-F-P → fleet fails
Fleet 5: 82-84-39-31-83-14-34 → P-P-P-P-P-F-P → fleet fails (c) 4 / 5 = 0.80

3.37 (a) 13. (b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9 am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

3.39 (a) E(X) = 3.59. SD(X) = 9.64. (b) E(X) = -1.41. SD(X) = 9.64. (c) No, the expected net profit is negative, so on average you expect to lose money.

3.41 5% increase in value.
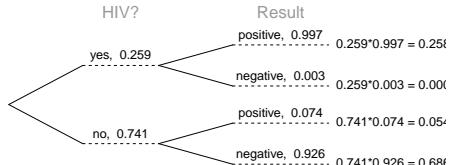
3.43 E = -0.0526. SD = 0.9986.

3.45 Approximate answers are OK.
(a) $(29 + 32)/144 = 0.42$. (b) $21/144 = 0.15$. (c) $(26 + 12 + 15)/144 = 0.37$.

3.47 (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) Invalid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

3.49 0.8247.
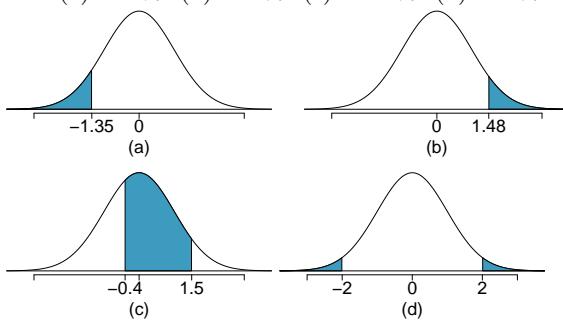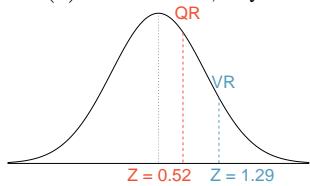
3.51 (a) E = \$3.90. SD = \$0.34. (b) E = \$27.30. SD = \$0.89.

4 Distributions of random variables

4.1 (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.

(a)

(b)

(c)

(d)

4.3 (a) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$.

QR

VR

$Z = 0.52$ $Z = 1.29$

(b) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section. (c) She did better on the Verbal Reasoning section since her Z-score on that section was higher. (d) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (e) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ did better than her on QR. (f) We cannot compare the raw scores since they are on different scales. Comparing her percentile scores is more appropriate when comparing her performance to others. (g) Answer to part (b) would not change as Z-scores can be calculated for distributions that are not normal. However, we could not answer parts (c)-(e) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

4.5 (a) $Z = 0.84$, which corresponds to approximately 160 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

4.7 (a) $Z = 1.2 \rightarrow 0.1151$. (b) $Z = -1.28 \rightarrow 70.6°$F or colder.

4.9 (a) $Z = 1.08 \rightarrow 0.1401$. (b) The answers are very close because only the units were changed. (The only reason why they are a little different is because 28°C is 82.4°F, not precisely 83°F.) (c) Since $IQR = Q3 - Q1$, we first need to find Q3 and Q1 and take the difference between the two. Remember that Q3 is the 75th and Q1 is the 25th percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26.86 - 23.13 = 3.73.

4.11 14/20 = 70% are within 1 SD. Within 2 SD: 19/20 = 95%. Within 3 SD: 20/20 = 100%. They follow this rule closely.

4.13 (a) Let X represent the amount of lemonade in the pitcher, Y represent the amount of lemonade in a glass, and W represent the amount left over after. Then, $\mu_W = E(X - Y) = 64 - 12 = 52$ (b) $\sigma_W = \sqrt{SD(X)^2 + SD(Y)^2} = \sqrt{1.732^2 + 1^2} \approx \sqrt{4} = 2$ (c) $P(W > 50) = P\left(Z > \frac{50-52}{2}\right) = P(Z > -1) = 1 - 0.1587 = 0.8413$
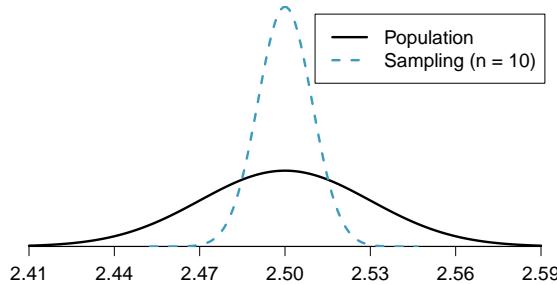
4.15 (a) The combined scores follow a normal distribution with $\mu_{combined} = 304$ and $\sigma_{combined} = 10.38$. Then, P(combined score > 320) is approximately 0.06. (b) Z=1.28 (using calculator or table). Then we set $1.28 = \frac{x-304}{10.38}$ and find $x \approx 317$.

4.17 (a) The distribution is unimodal and strongly right skewed with a median between 5 and 10 years old. Ages range from 0 to slightly over 50 years old, and the middle 50% of the distribution is roughly between 5 and 15 years old. There are potential outliers on the higher end. (b) When the sample size is small, the sampling distribution is right skewed, just like the population distribution. As the sample size increases, the sampling distribution gets more unimodal, symmetric, and approaches normality. The variability also decreases. This is consistent with the Central Limit Theorem.

4.19 (a) Right skewed. There is a long tail on the higher end of the distribution but a much shorter tail on the lower end. (b) Less than, as the median would be less than the mean in a right skewed distribution. (c) We should not. (d) Even though the population distribution is not normal, the conditions for inference are reasonably satisfied, with the possible exception of skew. If the skew isn't very strong (we should ask to see the data), then we can use the Central Limit Theorem to estimate this probability. For now, we'll assume the skew isn't very strong, though the description suggests it is at least moderate to strong. Use $N(1.3, SD_{\bar{x}} = 0.3/\sqrt{60})$: $Z = 2.58 \rightarrow 0.0049$. (e) It would decrease it by a factor of $\sqrt{2}$.

4.21 The centers are the same in each plot, and each data set is from a nearly normal distribution (see Section 7.1.1), though the histograms may not look very normal since each represents only 100 data points. The only way to tell which plot corresponds to which scenario is to examine the variability of each distribution. Plot B is the most variable, followed by Plot A, then Plot C. This means Plot B will correspond to the original data, Plot A to the sample means with size 5, and Plot C to the sample means with size 25.

4.23 (a) $Z = -3.33 \to 0.0004$. (b) The population SD is known and the data are nearly normal, so the sample mean will be nearly normal with distribution $N(\mu, \sigma/\sqrt{n})$, i.e. $N(2.5, 0.0095)$. (c) $Z = -10.54 \to \approx 0$. (d) See below:

(e) We could not estimate (a) without a nearly normal population distribution. We also could not estimate (c) since the sample size is not sufficient to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

4.25 (a) We cannot use the normal model for this calculation, but we can use the histogram. About 500 songs are shown to be longer than 5 minutes, so the probability is about $500/3000 = 0.167$. (b) Two different answers are reasonable. [Option 1] Since the population distribution is only slightly skewed to the right, even a small sample size will yield a nearly normal sampling distribution. We also know that the songs are sampled randomly and the sample size is less than 10% of the population, so the length of one song in the sample is independent of another. We are looking for the probability that the total length of 15 songs is more than 60 minutes, which means that the average song should last at least $60/15 = 4$ minutes. Using $SD_{\bar{x}} = 1.63/\sqrt{15}$, $Z = 1.31 \to 0.0951$. [Option 2] Since the population distribution is not normal, a small sample size may not be sufficient to yield a nearly normal sampling distribution. Therefore, we cannot estimate the probability using the tools we have learned so far. (c) We can now be confident that the conditions are satisfied. $Z = 0.92 \to 0.1788$.

4.27 (a) $SD_{\bar{x}} = \frac{25}{\sqrt{75}} = 2.89$. (b) $Z = 1.73$, which indicates that the two values are not unusually distant from each other when accounting for the uncertainty in John's point estimate.

4.29 (a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified. (b) No. There are six events under consideration. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6 and not rolling a 6,

though specifying such details would be necessary.

4.31 (a) $0.875^2 \times 0.125 = 0.096$. (b) $\mu = 8$, $\sigma = 7.48$.

4.33 If p is the probability of a success, then the mean of a Bernoulli random variable X is given by
$\mu = E[X] = P(X = 0) \times 0 + P(X = 1) \times 1$
$= (1 - p) \times 0 + p \times 1 = 0 + p = p$

4.35 (a) $\mu = 35$, $\sigma = 3.24$. (b) Yes. $Z = 3.09$. Since 45 is more than 2 standard deviations from the mean, it would be considered unusual. Note that the normal model is not required to apply this rule of thumb. (c) Using a normal model: 0.0010. This does indeed appear to be an unusual observation. If using a normal model with a 0.5 correction, the probability would be calculated as 0.0017.

4.37 (a) $1 - 0.75^3 = 0.5781$. (b) 0.1406. (c) 0.4219. (d) $1 - 0.25^3 = 0.9844$.

4.39 (a) Each observation in each of the distributions represents the sample proportion ($\hat{p}$) from samples of size $n = 20$, $n = 100$, and $n = 500$, respectively. (b) The centers for all three distributions are at 0.95, the true population parameter. When n is small, the distribution is skewed to the left and not smooth. As n increases, the variability of the distribution (standard deviation) decreases, and the shape of the distribution becomes more unimodal and symmetric.

4.41 (a) $SD_{\hat{p}} = \sqrt{p(1-p)/n} = 0.0707$. This describes the typical distance that the sample proportion will deviate from the true proportion, $p = 0.5$. (b) $\hat{p}$ approximately follows $N(0.5, 0.0707)$. $Z = (0.55 - 0.50)/0.0707 \approx 0.71$. This corresponds to an upper tail of about 0.2389. That is, $P(\hat{p} > 0.55) \approx 0.24$.

4.43 (a) First we need to check that the necessary conditions are met. There are $200 \times 0.08 = 16$ expected successes and $200 \times (1 - 0.08) = 184$ expected failures, therefore the success-failure condition is met. Then the binomial distribution can be approximated by $N(\mu = 16, \sigma = 3.84)$. $P(X < 12) = P(Z < -1.04) = 0.1492$. (b) Since the success-failure condition is met the sampling distribution of $\hat{p} \sim N(\mu = 0.08, \sigma = 0.0192)$. $P(\hat{p} < 0.06) = P(Z < -1.04) = 0.1492$. (c) As expected, the two answers are the same.

4.45 0 wins (-\$3): 0.1458. 1 win (-\$1): 0.3936. 2 wins (+\$1): 0.3543. 3 wins (+\$3): 0.1063.

4.47 Want to find the probability that there will be 1,786 or more enrollees. Using the normal approximation, with $\mu = np = 2,500 \times 0.7 = 1750$ and $\sigma = \sqrt{np(1-p)} = \sqrt{2,500 \times 0.7 \times 0.3} \approx 23$, $Z = 1.61$, and $P(Z > 1.61) = 0.0537$. With a 0.5 correction: 0.0559.

4.49 (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \to \sigma = \$223.88$.

4.51 (a) $(1-0.471)^2 \times 0.471 = 0.1318$. (b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$, $\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$, $\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials before a success and the standard deviation of the waiting time are higher.

4.53 $Z = 1.56$, $P(Z > 1.56) = 0.0594$, i.e. 6%.

4.55 (a) $Z = 0.73$, $P(Z > 0.73) = 0.2327$. (b) If you are bidding on only one auction and set a low maximum bid price, someone will probably outbid you. If you set a high maximum bid price, you may win the auction but pay more than is necessary. If bidding on more than one auction, and you set your maximum bid price very low, you probably won't win any of the auctions. However, if the maximum bid price is even modestly high, you are likely to win multiple auctions. (c) An answer roughly equal to the 10th percentile would be reasonable. Regrettably, no percentile cutoff point guarantees beyond any possible event that you win at least one auction. However, you may pick a higher percentile if you want to be more sure of winning an auction. (d) Answers will vary a

little but should correspond to the answer in part (c). We use the 10th percentile: $Z = -1.28 \to \$69.80$.

4.57 (a) $Z = 3.5$, upper tail is 0.0002. (More precise value: 0.000233, but we'll use 0.0002 for the calculations here.) (b) $0.0002 \times 2000 = 0.4$. We would expect about 0.4 10 year olds who are 76 inches or taller to show up. (c) $\binom{2000}{0}(0.0002)^0(1-0.0002)^{2000} = 0.67029$. (d) $\frac{0.4^0 \times e^{-0.4}}{0!} = \frac{1 \times e^{-0.4}}{1} = 0.67032$.

4.59 This is the same as checking that the average bag weight of the 10 bags is greater than 46 lbs. $SD_{\bar{x}} = \frac{3.2}{\sqrt{10}} = 1.012$; $z = \frac{46-45}{1.012} = 0.988$; $P(z > 0.988) = 0.162 = 16.2\%$.

4.61 First we need to check that the necessary conditions are met. There are $100 \times 0.389 = 38.9$ expected successes and $100 \times (1-0.389) = 61.1$ expected failures, therefore the success-failure condition is met. Calculate using either (1) the normal approximation to the binomial distribution or (2) the sampling distribution of $\hat{p}$. (1) The binomial distribution can be approximated by $N(\mu = 0.389, \sigma = 4.88)$. $P(X \geq 35) = P(Z > -0.80) = 1 - 0.2119 = 0.7881$. (2) The sampling distribution of $\hat{p} \sim N(\mu = 0.389, \sigma = 0.0488)$. $P(\hat{p} > 0.35) = P(Z > -0.8) = 0.7881$.

5 Foundations for inference

5.1 (a) Mean. Each student reports a numerical value: a number of hours. (b) Mean. Each student reports a number, which is a percentage, and we can average over these percentages. (c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion. (d) Mean. Each student reports a number, which is a percentage like in part (b). (e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion.

5.3 (a) The sample is from all computer chips manufactured at the factory during the week of production. We might be tempted to generalize the population to represent all weeks, but we should exercise caution here since the rate of defects may change over time. (b) The fraction of computer chips manufactured at the factory during the week of production that had defects. (c) Estimate the parameter using the data: $\hat{p} = \frac{27}{212} = 0.127$. (d) *Standard error* (or *SE*). (e) Compute the *SE* using $\hat{p} = 0.127$ in place of p: $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023$. (f) The standard error is the standard deviation of $\hat{p}$. A value of 0.10 would be about one standard error away from the observed value, which would not represent a very uncommon deviation. (Usually beyond about 2 stan-

dard errors is a good rule of thumb.) The engineer should not be surprised. (g) Recomputed standard error using $p = 0.1$: $SE = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$. This value isn't very different, which is typical when the standard error is computed using relatively similar proportions (and even sometimes when those proportions are quite different!).

5.5 (a) Sampling distribution. (b) If the population proportion is in the 5-30% range, the success-failure condition would be satisfied and the sampling distribution would be symmetric. (c) We use the formula for the standard error: $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$. (d) Standard error. (e) The distribution will tend to be more variable when we have fewer observations per sample.

5.7 Recall that the general formula is *point estimate* $\pm z^\star \times SE$. First, identify the three different values. The point estimate is 45%, $z^\star = 1.96$ for a 95% confidence level, and $SE = 1.2\%$. Then, plug the values into the formula: $45\% \pm 1.96 \times 1.2\% \to (42.6\%, 47.4\%)$ We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%.

5.9 (a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (b) True. Notice that the description focuses on the true population value. (c) True. If we examine the 95% confidence interval computed in Exercise 5.9, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5. (d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.

5.11 (a) False. Inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval. (b) True. (c) False. The confidence interval is not about a sample mean. (d) False. To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake. (e) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval. (f) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample $2^2 = 4$ times the number of people in the initial sample.

5.13 (a) $H_0 : p = 0.5$ (Neither a majority nor minority of students' grades improved) $H_A : p \neq 0.5$ (Either a majority or a minority of students' grades improved)
(b) $H_0 : \mu = 15$ (The average amount of company time each employee spends not working is 15 minutes for March Madness.) $H_A : \mu \neq 15$ (The average amount of company time each employee spends not working is different than 15 minutes for March Madness.)

5.15 (1) The hypotheses should be about the population proportion (p), not the sample proportion. (2) The null hypothesis should have an equal sign. (3) The alternative hypothesis should have a not-equals sign, and (4) it should reference the null value, $p_0 = 0.6$, not the observed sample proportion. The correct way to set up these hypotheses is: $H_0 : p = 0.6$ and $H_A : p \neq 0.6$.

5.17 (a) This claim is reasonable, since the entire interval lies above 50%. (b) The value of 70% lies out-

side of the interval, so we have convincing evidence that the researcher's conjecture is wrong. (c) A 90% confidence interval will be narrower than a 95% confidence interval. Even without calculating the interval, we can tell that 70% would not fall in the interval, and we would reject the researcher's conjecture based on a 90% confidence level as well.

5.19 (i) Set up hypotheses. H_0: $p = 0.5$, H_A: $p \neq 0.5$. We will use a significance level of $\alpha = 0.05$. (ii) Check conditions: simple random sample gets us independence, and the success-failure conditions is satisfied since $0.42 \times 1000 = 420$ and $(1-0.42) \times 1000 = 580$ are both at least 10. (iii) Next, we calculate: $SE = \sqrt{0.5(1-0.5)/1000} = 0.016$. $Z = \frac{0.42-0.5}{0.016} = -5$, which has a one-tail area of about 0.0000003, so the p-value is twice this one-tail area at 0.0000006. (iv) Make a conclusion: Because the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that the fraction of US adults who believe raising the minimum wage will help the economy is not 50%. Because the observed value is less than 50% and we have rejected the null hypothesis, we can conclude that this belief is held by fewer than 50% of US adults. (For reference, the survey also explores support for changing the minimum wage, which is a different question than if it will help the economy.)

5.21 If the p-value is 0.05, this means the test statistic would be either $Z = -1.96$ or $Z = 1.96$. We'll show the calculations for $Z = 1.96$. Standard error: $SE = \sqrt{0.3(1-0.3)/90} = 0.048$. Finally, set up the test statistic formula and solve for $\hat{p}$: $1.96 = \frac{\hat{p}-0.3}{0.048} \rightarrow \hat{p} = 0.394$ Alternatively, if $Z = -1.96$ was used: $\hat{p} = 0.206$.

5.23 (a) H_0: Anti-depressants do not affect the symptoms of Fibromyalgia. H_A: Anti-depressants do affect the symptoms of Fibromyalgia (either helping or harming). (b) Concluding that anti-depressants either help or worsen Fibromyalgia symptoms when they actually do neither. (c) Concluding that anti-depressants do not affect Fibromyalgia symptoms when they actually do.

5.25 (a) We are 95% confident that Americans spend an average of 1.38 to 1.92 hours per day relaxing or pursuing activities they enjoy. (b) Their confidence level must be higher as the width of the confidence interval increases as the confidence level increases. (c) The new margin of error will be smaller, since as the sample size increases, the standard error decreases, which will decrease the margin of error.

5.27 (a) H_0: The restaurant meets food safety and sanitation regulations. H_A: The restaurant does not meet food safety and sanitation regulations. (b) The food safety inspector concludes that the restaurant does not meet food safety and sanitation regulations and shuts down the restaurant when the restaurant is actually safe. (c) The food safety inspector concludes that the restaurant meets food safety and sanitation regulations and the restaurant stays open when the restaurant is actually not safe. (d) A Type 1 Error may be more problematic for the restaurant owner since his restaurant gets shut down even though it meets the food safety and sanitation regulations. (e) A Type 2 Error may be more problematic for diners since the restaurant deemed safe by the inspector is actually not. (f) Strong evidence. Diners would rather a restaurant that meet the regulations get shut down than a restaurant that doesn't meet the regulations not get shut down.

5.29 (a) H_0 : $p_{unemp} = p_{underemp}$: The proportions of unemployed and underemployed people who are having relationship problems are equal. H_A : $p_{unemp} \neq punderemp$: The proportions of unemployed and underemployed people who are having relationship problems are different. (b) If in fact the two population proportions are equal, the probability of observing at least a 2% difference between the sample proportions is approximately 0.35. Since this is a high probability we fail to reject the null hypothesis. The data do not provide convincing evidence that the proportion of of unemployed and underemployed people who are having relationship problems are different.

5.31 Because 130 is inside the confidence interval, we do not have convincing evidence that the true average is any different than what the nutrition label suggests.

5.33 True. If the sample size gets ever larger, then the standard error will become ever smaller. Eventually, when the sample size is large enough and the standard error is tiny, we can find statistically sig-

nificant yet very small differences between the null value and point estimate (assuming they are not exactly equal).

5.35 (a) In effect, we're checking whether men are paid more than women (or vice-versa), and we'd expect these outcomes with either chance under the null hypothesis:

$$H_0 : p = 0.5 \qquad H_A : p \neq 0.5$$

We'll use p to represent the fraction of cases where men are paid more than women.
(b) Below is the completion of the hypothesis test.

- There isn't a good way to check independence here since the jobs are not a simple random sample. However, independence doesn't seem unreasonable, since the individuals in each job are different from each other. The success-failure condition is met since we check it using the null proportion: $p_0 n = (1 - p_0)n = 10.5$ is greater than 10.

- We can compute the sample proportion, SE, and test statistic:

$$\hat{p} = 19/21 = 0.905$$
$$SE = \sqrt{\frac{0.5 \times (1 - 0.5)}{21}} = 0.109$$
$$Z = \frac{0.905 - 0.5}{0.109} = 3.72$$

  The test statistic Z corresponds to an upper tail area of about 0.0001, so the p-value is 2 times this value: 0.0002.

- Because the p-value is smaller than 0.05, we reject the notion that all these gender pay disparities are due to chance. Because we observe that men are paid more in a higher proportion of cases and we have rejected H_0, we can conclude that men are being paid higher amounts in ways not explainable by chance alone.

If you're curious for more info around this topic, including a discussion about adjusting for additional factors that affect pay, please see the following video by Healthcare Triage: youtu.be/aVhgKSULNQA.

6 Inference for categorical data

6.1 (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.

6.3 (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.

6.5 (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 82% $\pm$ 2%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.

6.7 With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^{\star}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times0.44}{600}} = 0.0397 \approx 4\%$

6.9 (a) No. The sample only represents students who took the SAT, and this was also an online survey. (b) (0.5289, 0.5711). We are 90% confident that 53% to 57% of high school seniors who took the SAT are fairly certain that they will participate in a study abroad program in college. (c) 90% of such random samples would produce a 90% confidence interval that includes the true proportion. (d) Yes. The interval lies entirely above 50%.

6.11 (a) We want to check for a majority (or minority), so we use the following hypotheses:

$$H_0 : p = 0.5 \qquad\qquad H_A : p \neq 0.5$$

We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents.
Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617 \times (1-0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test).
Therefore, we can model $\hat{p}$ using a normal distribu-

tion with a standard error of

$$SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$$

(We use the null proportion $p_0 = 0.5$ to compute the standard error for a one-proportion hypothesis test.) Next, we compute the test statistic:

$$Z = \frac{0.55 - 0.5}{0.02} = 2.5$$

This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$.
Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit.
(b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g. a 99% confidence level and a $\alpha = 0.05$ significance level), then this is no longer generally true.

6.13 (a) $H_0 : p = 0.5$. $H_A : p \neq 0.5$. Independence (random sample) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ the one tail area is 0.0018, so the p-value is 0.0036. Since the p-value < 0.05, we reject the null hypothesis. Since we rejected H_0 and the point estimate suggests people are better than random guessing, we can conclude the rate of correctly identifying a soda for these people is significantly better than just by random guessing. (b) If in fact people cannot tell the difference between diet and regular soda and they were randomly guessing, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly (or 53 or more identify a soda incorrectly) would be 0.0036.

6.15 Since a sample proportion ($\hat{p} = 0.55$) is available, we use this for the sample size calculations. The margin of error for a 90% confidence interval is $1.65 \times SE = 1.65 \times \sqrt{\frac{p(1-p)}{n}}$. We want this to be less than 0.01, where we use $\hat{p}$ in place of p:

$$1.65 \times \sqrt{\frac{0.55(1 - 0.55)}{n}} \leq 0.01$$
$$1.65^2 \frac{0.55(1 - 0.55)}{0.01^2} \leq n$$

From this, we get that n must be at least 6739.

6.17 This is not a randomized experiment, and it is unclear whether people would be affected by the behavior of their peers. That is, independence may not hold. Additionally, there are only 5 interventions under the provocative scenario, so the success-failure condition does not hold. Even if we consider a hypothesis test where we pool the proportions, the success-failure condition will not be satisfied. Since one condition is questionable and the other is not satisfied, the difference in sample proportions will not follow a nearly normal distribution.

6.19 (a) False. The entire confidence interval is above 0. (b) True. (c) True. (d) True. (e) False. It is simply the negated and reordered values: (-0.06,-0.02).

6.21 (a) Standard error:

$$SE = \sqrt{\frac{0.79(1 - 0.79)}{347} + \frac{0.55(1 - 0.55)}{617}} = 0.03$$

Using $z^\star = 1.96$, we get:

$$0.79 - 0.55 \pm 1.96 \times 0.03 \rightarrow (0.181, 0.299)$$

We are 95% confident that the proportion of Democrats who support the plan is 18.1% to 29.9% higher than the proportion of Independents who support the plan. (b) True.

6.23 (a) College grads: 23.7%. Non-college grads: 33.7%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college graduates who responded "do not know". $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 235/827 = 0.284$), is also satisfied. $Z = -3.18 \rightarrow$ p-value = 0.0014. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates. The data also indicate that fewer college grads say they "do not know" than non-college grads (i.e. the data indicate the direction after we reject H_0).

6.25 (a) College grads: 35.2%. Non-college grads: 33.9%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college grads who support offshore drilling. $H_0 : p_{CG} = p_{NCG}$.

$H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 286/827 = 0.346$), is also satisfied. $Z = 0.39 \rightarrow$ p-value = 0.6966. Since the p-value $> \alpha$ (0.05), we fail to reject H_0. The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support off-shore drilling in California.

6.27 Subscript $_C$ means control group. Subscript $_T$ means truck drivers. $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 70/495 = 0.141$). $Z = -1.65 \rightarrow$ p-value = 0.0989. Since the p-value is high (default to alpha = 0.05), we fail to reject H_0. The data do not provide strong evidence that the rates of sleep deprivation are different for non-transportation workers and truck drivers.

6.29 (a) Summary of the study:

		Virol. failure		
		Yes	No	Total
Treatment	Nevaripine	26	94	120
	Lopinavir	10	110	120
	Total	36	204	240

(b) $H_0 : p_N = p_L$. There is no difference in virologic failure rates between the Nevaripine and Lopinavir groups. $H_A : p_N \neq p_L$. There is some difference in virologic failure rates between the Nevaripine and Lopinavir groups. (c) Random assignment was used, so the observations in each group are independent. If the patients in the study are representative of those in the general population (something impossible to check with the given information), then we can also confidently generalize the findings to the population. The success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 36/240 = 0.15$), is satisfied. $Z = 2.89 \rightarrow$ p-value = 0.0039. Since the p-value is low, we reject H_0. There is strong evidence of a difference in virologic failure rates between the Nevaripine and Lopinavir groups. Treatment and virologic failure do not appear to be independent.

6.31 (a) False. The chi-square distribution has one parameter called degrees of freedom. (b) True. (c) True. (d) False. As the degrees of freedom increases, the shape of the chi-square distribution becomes more symmetric.

6.33 (a) H_0: The distribution of the format of the book used by the students follows the professor's predictions. H_A: The distribution of the format of the book used by the students does not follow the professor's predictions. (b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$. $E_{print} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$. (c) Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, p-value $= 0.313$. (e) Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

6.35 (a) Two-way table:

Treatment	Quit Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total) \times (col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value.
(b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total) \times (col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

6.37 H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California. H_A: Opinions regarding the drilling for oil and natural gas off the coast of California has an association with earning a college degree.

$$E_{row\ 1,col\ 1} = 151.5 \qquad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \qquad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \qquad E_{row\ 3,col\ 2} = 110.5$$

Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence between observations is reasonable. Sample size: All expected counts are at least 5. $\chi^2 = 11.47$, $df = 2 \rightarrow$ p-value $= 0.003$. Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is an association between support for offshore drilling and having a college degree.

6.39 No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

6.41 (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

6.43 (a) Independence is satisfied (random sample), as is the success-failure condition (40 smokers, 160 non-smokers). The 95% CI: (0.145, 0.255). We are 95% confident that 14.5% to 25.5% of all students at this university smoke. (b) We want $z^\star SE$ to be no larger than 0.02 for a 95% confidence level. We use $z^\star = 1.96$ and plug in the point estimate $\hat{p} = 0.2$ within the SE formula: $1.96\sqrt{0.2(1 - 0.2)/n} \leq 0.02$. The sample size n should be at least 1,537.

6.45 (a) Proportion of graduates from this university who found a job within one year of graduating. $\hat{p} = 348/400 = 0.87$. (b) This is a random sample, so the observations are independent. Success-failure condition is satisfied: 348 successes, 52 failures, both well above 10. (c) (0.8371, 0.9029). We are 95% confident that approximately 84% to 90% of graduates from this university found a job within one year of completing their undergraduate degree. (d) 95% of such random samples would produce a 95% confidence interval that includes the true proportion of students at this university who found a job within one year of graduating from college. (e) (0.8267, 0.9133). Similar interpretation as before. (f) 99% CI is wider, as we are more confident that the true proportion is within the interval and so need to cover a wider range.

6.47 Use a chi-squared goodness of fit test. H_0: Each option is equally likely. H_A: Some options are preferred over others. Total sample size: 99. Expected counts: $(1/3) * 99 = 33$ for each option. These are all above 5, so conditions are satisfied. $df = 3 - 1 = 2$ and $\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52 \rightarrow$ p-value $= 0.023$. Since the p-value is less than 5%, we reject H_0. The data provide convincing evidence that some options are preferred over others.

6.49 (a) $H_0 : p = 0.38$. $H_A : p \neq 0.38$. Independence (random sample) and the success-failure condition are satisfied. $Z = -20.5 \rightarrow$ p-value ≈ 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%, and the data indicate that the proportion is lower in the US. (b) If in fact 38% of Americans used their cell phones as a primary access point to the internet, the probability of obtaining a random sample of 2,254 Americans where 17% or less or 59% or more use their only their cell phones to access the internet would be approximately 0. (c) (0.1545, 0.1855). We are 95% confident that approximately 15.5% to 18.6% of all Americans primarily use their cell phones to browse the internet.

7 Inference for numerical data

7.1 (a) $df = 6 - 1 = 5$, $t_5^\star = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 20$, $t_{20}^\star = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t_{28}^\star = 2.05$. (d) $df = 11$, $t_{11}^\star = 3.11$.

7.3 (a) 0.085, do not reject H_0. (b) 0.003, reject H_0. (c) 0.438, do not reject H_0. (d) 0.042, reject H_0.

7.5 The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t_{35}^\star = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.

7.7 (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.) (b) Independence: The sample is random. The min/max suggest there are no concerning outliers. $T = -1.75$. $df = 25 - 1 = 24$. (c) p-value = 0.093. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hours per night or less (or 8.27 hours or more) is 0.093. (d) Since p-value > 0.05, do not reject H_0. The data do not provide strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) Yes, since we did not rejected H_0.

7.9 T is either -2.09 or 2.09. Then $\bar{x}$ is one of the following:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

7.11 (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu < 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{30} = 0.402$. The test statistic is $T = (4.6 - 5)/SE = -0.995$. $df = 30 - 1 = 29$. The p-value is about 0.164, which is bigger than $\alpha = 0.05$ and we do not reject H_0. That is, we do not have sufficiently strong evidence to reject Georgianna's claim that the average is (at least) 5 years.
(b) Using $SE = 0.402$ and $t_{df=29}^\star = 2.045$, the confidence interval is (3.78, 5.42). We are 95% confident that the average number of years a child takes piano lessons in this city is between 3.78 and 5.42 years.

(c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

7.13 Assuming the population standard deviation is known, the margin of error will be $1.96 \times 100/\sqrt{n}$. We want this value to be less than 10, which leads to $n \geq 384.16$, meaning we need a sample size of at least 385 (round up for sample size calculations!).

7.15 Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point.

7.17 (a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent. (b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group. (c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent. (d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

7.19 (a) For each observation in one data set, there is exactly one specially corresponding observation in the other data set for the same geographic location. The data are paired. Let $diff = 2018 - 1948$. (b) $H_0 : \mu_{\text{diff}} = 0$ (On average, the number of days exceeding 90°F in 1948 and 2018 for NOAA stations was the same.) $H_A : \mu_{\text{diff}} > 0$ (On average, there were more days exceeding 90°F in 2018 than 1948 for NOAA stations.) (c) Locations were randomly sampled, so independence is reasonable. The sample size is at least 30, so we're just looking for particularly extreme outliers: none are present (the observation off left in the histogram would be considered a clear outlier, but not a particularly extreme one). Therefore, the conditions are satisfied. (d) $SE = 17.2/\sqrt{197} = 1.23$. $T = \frac{2.9 - 0}{1.23} = 2.36$ with degrees of freedom $df = 197 - 1 = 196$. This leads to a p-value of about 0.019. (e) Since the p-value is less than 0.05, we reject H_0. The data provide strong evidence that NOAA stations observed more 90°F days in 2018 than in 1948. (f) Type 1 Error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe a decrease, but the sample we took just so happened to make it appear that this was the case.

7.21 (a) $SE = 1.23$ and $z^{\star} = 1.65$. $2.9 \pm 1.65 \times 1.23 \to (0.87, 4.93)$.
(b) We are 90% confident that there was an increase of 0.87 to 4.93 in the average number of days that hit 90°F in 2018 relative to 1948 for NOAA stations.
(c) Yes, since the interval lies entirely above 0.

7.23 (a) These data are paired. For example, the Friday the 13th in say, September 1991, would probably be more similar to the Friday the 6th in September 1991 than to Friday the 6th in another month or year.
(b) Let $\mu_{diff} = \mu_{sixth} - \mu_{thirteenth}$. $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$.
(c) Independence: The months selected are not random. However, if we think these dates are roughly equivalent to a simple random sample of all such Friday 6th/13th date pairs, then independence is reasonable. To proceed, we must make this strong assumption, though we should note this assumption in any reported results. Normality: With fewer than 10 observations, we would need to see clear outliers to be concerned. There is a borderline outlier on the right of the histogram of the differences, so we would want to report this in formal analysis results.
(d) $T = 4.94$ for $df = 10 - 1 = 9 \to$ p-value = 0.001.
(e) Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of cars at the intersection is higher on Friday the 6th than on Friday the 13th. (We should exercise caution about generalizing the interpretation to all intersections or roads.)
(f) If the average number of cars passing the intersection actually was the same on Friday the 6th and 13th, then the probability that we would observe a test statistic so far from zero is less than 0.01.
(g) We might have made a Type 1 Error, i.e. incorrectly rejected the null hypothesis.

7.25 (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. p-value = 0.042. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of traffic accident related emergency room admissions are different between Friday the 6th and Friday the 13th. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6th relative to Friday the 13th.
(b) (-6.49, -0.17).
(c) This is an observational study, not an experiment, so we cannot so easily infer a causal intervention implied by this statement. It is true that there is a difference. However, for example, this does not mean that a responsible adult going out on Friday the 13th has a higher chance of harm than on any other night.

7.27 (a) Chicken fed linseed weighed an average of 218.75 grams while those fed horsebean weighed an average of 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken fed linseed.
(b) $H_0 : \mu_{ls} = \mu_{hb}$. $H_A : \mu_{ls} \neq \mu_{hb}$.
We leave the conditions to you to consider.
$T = 3.02$, $df = min(11, 9) = 9 \to$ p-value = 0.014. Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a significant difference between the average weights of chickens that were fed linseed and horsebean.
(c) Type 1 Error, since we rejected H_0.
(d) Yes, since p-value > 0.01, we would not have rejected H_0.

7.29 $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \to$ p-value = 0.007. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean (with weights from casein being higher). Since this is a randomized experiment, the observed difference can be attributed to the diet.

7.31 Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} \neq 0$: Treatment has an effect on P.D.T. scores, either positive or negative. Conditions: The subjects are randomly assigned to treatments, so independence within and between groups is satisfied. All three sample sizes are smaller than 30, so we look for clear outliers. There is a borderline outlier in the first treatment group. Since it is borderline, we will proceed, but we should report this caveat with any results. For all three groups: $df = 13$. $T_1 = 1.89 \to$ p-value = 0.081, $T_2 = 1.35 \to$ p-value = 0.200), $T_3 = -1.40 \to$ (p-value = 0.185). We do not reject the null hypothesis for any of these groups. As earlier noted, there is some uncertainty about if the method applied is reasonable for the first group.

7.33 $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \to$ p-value = 0.036. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average food consumption by the patients in the treatment and control groups are different. Furthermore, the data indicate patients in the distracted eating (treatment) group consume more food than patients in the control group.

7.35 False. While it is true that paired analysis requires equal sample sizes, only having the equal sample sizes isn't, on its own, sufficient for doing a paired test. Paired tests require that there be a special correspondence between each pair of observations in the two groups.

7.37 (a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution. (b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric. (c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = 18.2/\sqrt{45} = 2.713$. (d) The sample means will be more variable with the smaller sample size.

7.39 Independence: it is a random sample, so we can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclu-sive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported). The sample size is at least 30, and there are no particularly extreme outliers, so the normality condition is reasonable. 90% CI: (2.97, 3.43). We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

7.41 The hypotheses should be about the population mean (μ), not the sample mean. The null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. Correction:

$$H_0 : \mu = 10 \; hours$$
$$H_A : \mu \neq 10 \; hours$$

Because the change could go either way, we use a two-sided H_A.

8 Introduction to linear regression

8.1 (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

8.3 (a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.

8.5 (a) Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear. (b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam. Answers may vary.

8.7 (a) $r = -0.7 \rightarrow (4)$. (b) $r = 0.45 \rightarrow (3)$. (c) $r = 0.06 \rightarrow (1)$. (d) $r = 0.92 \rightarrow (2)$.

8.9 (a) The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot. (b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion. (c) Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

8.11 (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours. (c) Changing units doesn't affect correlation: $r = 0.636$.

8.13 (a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

8.15 In each part, we can write the woman's age as a linear function of the spouse's age.
(a) $age_W = age_S + 3$.
(b) $age_W = age_S - 2$.
(c) $age_W = 2 \times age_S$.
Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, e.g. 5 women aged 26, 27, 28, 29, and 30, then find the spouses ages for each women in each part and create a scatterplot.

8.17 Correlation: no units. Intercept: kg. Slope: kg/cm.

8.19 Over-estimate. Since the residual is calculated as *observed* − *predicted*, a negative residual means that the predicted value is higher than the observed value.

8.21 (a) There is a positive, very strong, linear association between the number of tourists and spending. (b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism. (d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

8.23 (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (b) b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time. b_0: When the distance traveled is 0

miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.

8.25 (a) $\widehat{murder} = -29.901 + 2.559 \times poverty\%$. (b) Expected murder rate in metropolitan areas with no poverty is -29. 901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559. (d) Poverty level explains 70.52% of the variability in murder rates in metropolitan areas. (e) $\sqrt{0.7052} = 0.8398$.

8.27 (a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope.
(b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point.
(c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.

8.29 (a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot. (b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

8.31 (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential.
(b) $\widehat{weight} = -105.0113 + 1.0176 \times height$.
Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds).
Intercept: People who are 0 centimeters tall are expected to weigh - 105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself.
(c) H_0: The true slope coefficient of height is zero ($\beta_1 = 0$).
H_A: The true slope coefficient of height is different than zero ($\beta_1 \neq 0$).
The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0.
(d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.

8.33 (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. The p-value, as reported in the table, is incredibly small and is smaller than 0.05, so we reject H_0. The data provide convincing evidence that womens' and spouses' heights are positively correlated.
(b) $\widehat{height}_S = 43.5755 + 0.2863 \times height_W$.
(c) Slope: For each additional inch in woman's height, the spouse's height is expected to be an additional 0.2863 inches, on average. Intercept: Women who are 0 inches tall are predicted to have spouses who are 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line.
(d) The slope is positive, so r must also be positive. $r = \sqrt{0.09} = 0.30$.
(e) 63.2612. Since R^2 is low, the prediction based on this regression model is not very reliable.
(f) No, we should avoid extrapolating.

8.35 (a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T_{18}^* = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million. (d) Yes, we rejected H_0 and the confidence interval does not include 0.

8.37 (a) The relationship is positive, non-linear, and somewhat strong. Due to the non-linear form of the relationship and the clear non-constant variance in the residuals, a linear model is not appropriate for modeling the relationship between year and price. (b) The logged model is a much better fit: the scatter plot shows a linear relationships and the residuals do not appear to have a pattern. (c) For each year increase in the year of the truck (for each year the truck is newer) we would expect the price of the truck to increase on average by a factor of $e^{0.137} \approx 1.15$, i.e. by 15%.

8.39 (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.

8.41 There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

8.43 (a) $r = -0.72 \rightarrow (2)$ (b) $r = 0.07 \rightarrow (4)$ (c) $r = 0.86 \rightarrow (1)$ (d) $r = 0.99 \rightarrow (3)$

8.45 (a) There is a weak-to-moderate, positive, linear association between height and volume. There also appears to be some non-constant variance since the volume of trees is more variable for taller trees. (b) There is a very strong, positive association between diameter and volume. The relationship may include slight curvature. (c) Since the relationship is stronger between volume and diameter, using diameter would be preferred. However, as mentioned in part (b), the relationship between volume and diameter may not be, and so we may benefit from a model that properly accounts for nonlinearity.

Appendix B

Data sets within the text

Each data set within the text is described in this appendix, and there is a corresponding page for each of these data sets at **openintro.org/data**. This page also includes additional data sets that can be used for honing your skills. Each data set has its own page with the following information:

- Description of each data set.

- Detailed overview of each data set's variables.

- CSV download.

- R object file download.

Over time we will also expand the information available on these pages.

Chapter 1: Data collection

1.1 `stent30`, `stent365` → The stent data is split across two data sets, one for the 0-30 day and one for the 0-365 day results.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
NY Times article: www.nytimes.com/2011/09/08/health/research/08stent.html.

1.2 `loan50`, `loans_full_schema` → This data comes from Lending Club (lendingclub.com), which provides a large set of data on the people who received loans through their platform. The data used in the textbook comes from a sample of the loans made in Q1 (Jan, Feb, March) 2018.

1.2 `county`, `county_complete` → These data come from several government sources. For those variables included in the county data set, only the most recent data is reported, as of what was available in late 2018. Data prior to 2011 is all from census.gov, where the specific Quick Facts page providing the data is no longer available. The more recent data comes from USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), and American Community Survey (census.gov/programs-surveys/acs).

1.4 The Nurses' Health Study was mentioned. For more information on this data set, see www.channing.harvard.edu/nhs

1.5 The study we had in mind when discussing the simple randomization (no blocking) study was Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

Chapter 2: Summarizing data

2.1 `loan50`, `county` → These data sets are described in the data for Chapter 1.

2.3 `loan50`, `county` → These data sets are described in the data for Chapter 1.

2.4 `malaria` → Lyke et al. 2017. PfSPZ vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. PNAS 114(10):2711-2716. www.pnas.org/content/114/10/2711

Chapter 3: Probability

3.1 `loan50`, `county` → These data sets are described in the data for Chapter 1.

3.1 `playing_cards` → A table describing the 52 cards in a standard deck.

3.2 `family_college` → A simulated data set based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

3.2 `smallpox` → Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

3.2 Mammogram screening, probabilities. → The probabilities reported were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

3.5 `stocks_18` → Monthly returns for Caterpillar, Exxon Mobil Corp, and Google for November 2015 to October 2018.

3.6 `fcid` → This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

Chapter 4: Distributions of random variables

4.1 SAT and ACT score distributions → The SAT score data comes from the 2018 distribution, which is provided at

reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf

The ACT score data is available at

act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf

We also acknowledge that the actual ACT score distribution is *not* nearly normal. However, since the topic is very accessible, we decided to keep the context and examples.

4.1 `possum` → The distribution parameters are based on a sample of possums from Australia and New Guinea. The original source of this data is as follows. Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala).* Australian Journal of Zoology 43: 449-458.

4.1 `male_heights_fcid` → This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

4.1 `nba_players_19` → Summary information from the NBA players for the 2018-2019 season. Data were retrieved from www.nba.com/players.

4.1 `poker` → Poker winnings (and losses) for 50 days by a professional poker player, which represents their first 50 days trying to play for a living. Anonymity has been requested by the player.

4.2 `run17`, `run17samp` → www.cherryblossom.org

Chapter 5: Foundations for inference

5.1 `pew_energy_2018` → The actual data has more observations than were referenced in this chapter. That is, we used a subsample since it helped smooth some of the examples to have a bit more variability. The `pew_energy_2018` data set represents the full data set for each of the different energy source questions, which covers solar, wind, offshore drilling, hydrolic fracturing, and nuclear energy. The statistics used to construct the data are from the following page:

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

5.2 `pew_energy_2018` → See the details for this data set above in the Section 5.1 data section.

5.2 `ebola_survey` → In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014. Poll ID NY141026 on maristpoll.marist.edu.

5.3 `pew_energy_2018` → See the details for this data set above in the Section 5.1 data section.

Chapter 6: Inference for categorical data

6.1 Nuclear energy → A Gallup poll of 1,019 adults in the US, conducted in March of 2016, found that 54% of respondents oppose nuclear energy. This was the first time since Gallup first asked the question in 1994 that a majority of respondents said they oppose nuclear energy.
https://news.gallup.com/poll/190064/first-time-majority-oppose-nuclear-energy.aspx

6.1 Supreme Court → The Gallup organization began measuring the public's view of the Supreme Court's job performance in 2000, and has measured it every year since then with the question: "Do you approve or disapprove of the way the Supreme Court is handling its job?". In 2018, the Gallup poll randomly sampled 1,033 adults in the U.S. and found that 53% of them approved.
https://news.gallup.com/poll/237269/supreme-court-approval-highest-2009.aspx

6.1 Life on other planets → A February 2018 Marist Poll reported: "Many Americans (68%) think there is intelligent life on other planets". The results were based on a random sample of 1,033 adults in the U.S.
http://maristpoll.marist.edu/212-are-americans-poised-for-an-alien-invasion

6.2 `cpr` → Böttiger et al. *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial.* The Lancet, 2001.

6.2 `fish_oil_18` → Manson JE, et al. 2018. Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer. NEJMoa1811403.

6.2 `mammogram` → Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

6.2 `drone_blades` → The quality control data set for quadcopter drone blades is a made-up data set for an example. We provide the simulated data in the **drone_blades** data set.

6.3 M&Ms → Starting at the end of 2016, Rick Wicklin, a statistician working at the statistical software company SAS, collected a sample of 712 candies, or about 1.5 pounds, and counted how many there were of each color.
https://qz.com/918008/the-color-distribution-of-mms-as-determined-by-a-phd-in-statistics

6.4 `ask` → Minson JA, Ruedy NE, Schweitzer ME. *There is such a thing as a stupid question: Question disclosure in strategic communication.*
opim.wharton.upenn.edu/DPlab/papers/workingPapers/
Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

6.4 `diabetes2` → Zeitler P, et al. 2012. A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes. N Engl J Med.

Chapter 7: Inference for numerical data

7.1 Risso's dolphins → Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins.

7.1 Croaker white fish → www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

7.1 run17, run17samp → www.cherryblossom.org

7.2 textbooks, ucla_textbooks_f18 → Data were collected by OpenIntro staff in 2010 and again in 2018. For the 2018 sample, we sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. The websites where information was retrieved: sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, and amazon.com.

7.3 Jennifer-John → Bertrand M, Mullainathan S. 2004. *Science faculty's subtle gender biases favor male students.* PNAS October 9, 2012 109 (41) 16474-16479. https://www.pnas.org/content/109/41/16474

7.3 resume → Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873

7.3 stem_cells → Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(05)67380-1/fulltext

Chapter 8: Introduction to linear regression

8.1 simulated_scatter → Fake data used for the first three plots. The perfect linear plot uses group 4 data, where group variable in the data set (Figure 8.1). The group of 3 imperfect linear plots use groups 1-3 (Figure 8.2). The sinusoidal curve uses group 5 data (Figure 8.3). The group of 3 scatterplots with residual plots use groups 6-8 (Figure 8.8). The correlation plots uses groups 9-19 data (Figures 8.9 and 8.10).

8.1 possum → This data is described in the data for Chapter 4.

8.2 elmhurst → These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

8.2 simulated_scatter → The plots for things that can go wrong uses groups 20-23 (Figure 8.22).

8.2 mariokart → Auction data from Ebay (ebay.com) for the game Mario Kart for the Nintendo Wii. This data set was collected in early October, 2009.

8.2 simulated_scatter → The plots for types of outliers uses groups 24-29 (Figure 8.19).

8.3 midterms_house → Data was retrieved from Wikipedia.

8.4 county, county_complete → This data is described in the data for Chapter 1.

Appendix C

Distribution tables
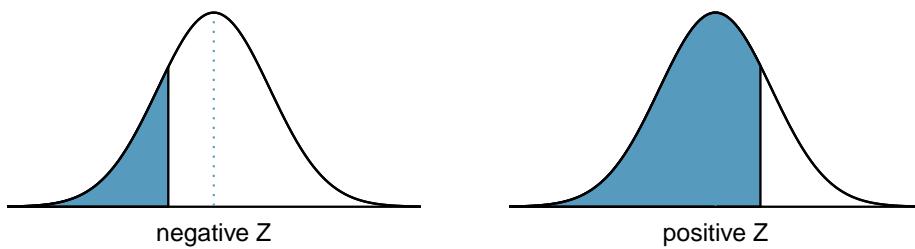
C.1 Random Number Table

Row	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
				Column				
1	44394	76100	85973	26853	07080	91603	00476	19681
2	61578	75037	54792	74216	31952	31235	31258	57886
3	18529	73285	95291	49606	67174	95905	33679	75811
4	81238	18321	71085	08284	39318	31434	26173	07440
5	11173	58878	25516	15058	48639	52723	95864	89673
6	96737	95194	14419	22202	92867	73525	94382	29927
7	63514	55066	65162	96016	91723	21160	24285	33264
8	35087	57036	10001	39424	50536	77380	45042	48180
9	00148	73933	49369	32403	53850	16291	93619	27557
10	28999	76232	32637	95697	63679	54506	11299	94294
11	37911	50834	10927	74075	26558	42311	36483	71820
12	33624	82379	03625	58336	27390	00586	06344	89625
13	93282	63059	10830	89432	26917	31555	51793	18718
14	57429	71933	80329	56521	97594	92651	14819	86546
15	65029	24328	06826	61448	54760	09351	73930	99564
16	14779	23173	97183	59835	69580	94653	55095	80666
17	52072	12187	35360	82925	44923	44532	18251	96991
18	76282	91849	17138	59554	35476	67007	02484	10122
19	46561	33015	04577	02178	32915	35912	48974	92985
20	70623	36097	48780	06921	60683	22461	36175	61281
21	03605	08541	17546	85790	48413	69382	89785	80206
22	46147	07603	92057	87609	52670	96255	96660	83167
23	09547	77804	95099	22158	53279	23161	72675	92804
24	12899	05005	86667	72331	09114	28187	97404	26750
25	21223	38353	56970	48965	58371	02697	61417	54746
26	35770	35697	32281	53514	10854	16778	56447	46965
27	04243	65817	81819	64381	83509	44316	56316	47742
28	56989	05587	79995	36598	02316	81627	50104	47720
29	53233	48698	59304	63566	25352	03322	29938	82306
30	20232	30909	77126	50041	96500	24033	77422	20150
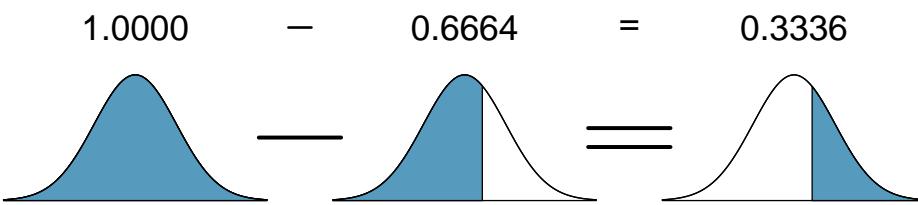
	Column							
Row	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
31	85882	59541	14275	15866	27467	60143	92033	22771
32	33055	24722	67250	27831	11114	84858	18231	85739
33	25994	89263	58632	08784	15774	27699	32181	52967
34	19103	95173	74832	68762	66983	16051	92092	72066
35	85452	51441	22086	47481	41880	98791	33532	10453
36	41523	69102	02604	00209	76159	99621	96573	59154
37	90311	26961	31394	43019	18521	54000	75983	46462
38	32669	76156	46877	86814	42652	74313	86128	95406
39	11155	77427	61695	16162	36682	27559	22972	20061
40	11132	57408	61007	97390	17122	53132	26189	21875
41	70967	21786	00053	32893	67681	81911	56693	15162
42	29013	28494	80802	38490	02808	54605	20490	19681
43	56896	71763	66787	16331	40798	22111	28907	81975
44	70658	25121	34292	99044	46390	86503	31601	82444
45	52392	67742	59495	16864	68170	95937	35545	84861
46	20741	67232	26971	27680	63048	95634	02828	22125
47	92549	56918	97969	92789	77949	70181	53477	68179
48	75084	02966	94937	04316	46782	03863	69626	24665
49	00063	53920	96953	16190	31447	63494	92765	38345
50	74807	86955	23214	84688	83291	12324	16325	81121
51	18186	13179	77206	57798	31333	69795	12667	31973
52	70135	99944	49928	79410	28233	83809	61091	47342
53	88043	90662	37325	41709	36888	28368	73822	10085
54	48258	76775	71829	85903	32278	03244	62429	11652
55	56399	28764	50930	63066	17125	47910	84486	85522
56	10879	94293	64826	35152	98776	96947	01132	84264
57	16355	60561	42182	17140	84048	32917	85483	68557
58	51190	14326	62013	10370	40045	64064	88484	08559
59	18762	84505	25892	90869	74228	53749	64947	95937
60	19150	85525	97008	81293	49517	41430	80339	20915
61	14294	08263	56326	04922	36882	89658	54217	90500
62	25913	60850	62974	06866	20111	38797	23664	21828
63	37278	97201	24337	49224	27299	28363	33961	59307
64	63837	80459	74548	93999	12775	81754	89349	23516
65	64551	65984	88299	61960	63880	41251	45278	80827
66	89928	97374	29847	35633	34776	65913	73208	25336
67	49108	79853	18853	40762	56218	98369	99315	99585
68	81354	69478	05333	57344	38877	02876	30826	59710
69	63308	04271	90756	98409	67880	15732	40799	70823
70	24368	25294	60570	71072	37576	71774	19587	38440
71	85617	05799	69763	50889	99515	36317	72949	27502
72	12557	39890	04807	49466	29763	72937	39541	64381
73	06607	02387	74363	75934	88791	35938	92553	92335
74	78809	28121	09576	60199	93428	86836	74682	29020
75	25180	36730	12967	18565	68906	90287	14317	94668

C.2 Normal Probability Table
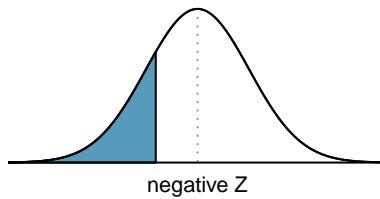
The area to the left of Z represents the percentile of the observation. The normal probability table always lists percentiles.

negative Z   positive Z

To find the area to the right, calculate 1 minus the area to the left.

1.0000   −   0.6664   =   0.3336
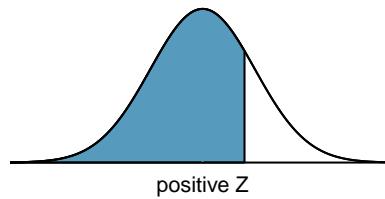
For additional details about working with the normal distribution and the normal probability table, see Section 4.1, which starts on page 199.

negative Z

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	-3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	-3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	-3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	-3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	-2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	-2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	-2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	-2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	-2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	-2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	-2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	-2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	-2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	-2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	-1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	-1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	-1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	-1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	-1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	-1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	-1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	-1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	-0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	-0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	-0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	-0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	-0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	-0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	-0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	-0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	-0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	-0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.

positive Z

Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

*For $Z \geq 3.50$, the probability is greater than or equal to 0.9998.

C.3 t Probability Table
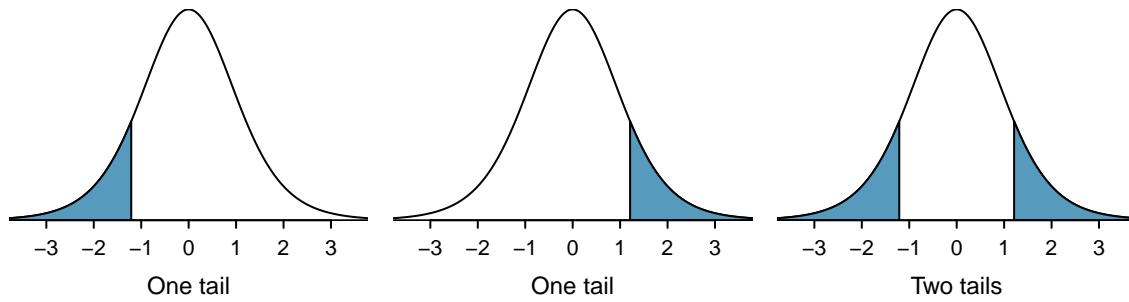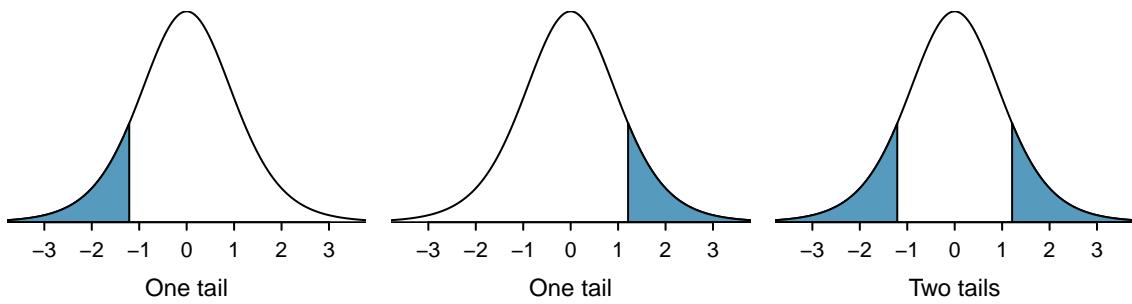
Figure C.1: Three t distributions.

	one tail	0.100	0.050	0.025	0.010	0.005
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77
	28	1.31	1.70	2.05	2.47	2.76
	29	1.31	1.70	2.05	2.46	2.76
	30	1.31	1.70	2.04	2.46	2.75
Confidence level C		80%	90%	95%	98%	99%

Figure C.2: Three t distributions.

one tail		0.100	0.050	0.025	0.010	0.005
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72
	36	1.31	1.69	2.03	2.43	2.72
	37	1.30	1.69	2.03	2.43	2.72
	38	1.30	1.69	2.02	2.43	2.71
	39	1.30	1.68	2.02	2.43	2.71
	40	1.30	1.68	2.02	2.42	2.70
	41	1.30	1.68	2.02	2.42	2.70
	42	1.30	1.68	2.02	2.42	2.70
	43	1.30	1.68	2.02	2.42	2.70
	44	1.30	1.68	2.02	2.41	2.69
	45	1.30	1.68	2.01	2.41	2.69
	46	1.30	1.68	2.01	2.41	2.69
	47	1.30	1.68	2.01	2.41	2.68
	48	1.30	1.68	2.01	2.41	2.68
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68
	60	1.30	1.67	2.00	2.39	2.66
	70	1.29	1.67	1.99	2.38	2.65
	80	1.29	1.66	1.99	2.37	2.64
	90	1.29	1.66	1.99	2.37	2.63
	100	1.29	1.66	1.98	2.36	2.63
	150	1.29	1.66	1.98	2.35	2.61
	200	1.29	1.65	1.97	2.35	2.60
	300	1.28	1.65	1.97	2.34	2.59
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.65	1.96	2.33	2.58
Confidence level C		80%	90%	95%	98%	99%

C.4   Chi-Square Probability Table
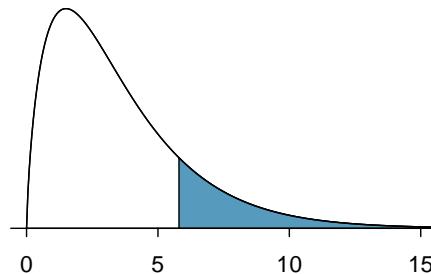
Figure C.3: Areas in the chi-square table always refer to the right tail.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66

Index

Appendix D

Calculator reference, Formulas, and Inference guide

D.1 Calculator reference

Instructions for the TI-83/84 and the Casio fx-9750GII, and their associated videos.

D.2 Formulas

Descriptive Statistics

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n}$$

$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

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

$$\hat{y} = a + bx$$

$$\bar{y} = a + b\bar{x}$$

$$b = r\frac{s_y}{s_x}$$

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

Probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\mu_X = E(X) = \sum x_i \cdot P(x_i)$$

$$\sigma_X = \sqrt{\sum(x_i - \mu_x)^2 \cdot P(x_i)}$$

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If X has a binomial distribution with parameters n and p, then:

$$P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$$

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Inferential Statistics

$$\text{standardized test statistic: } \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

$$\text{confidence interval: point estimate } \pm \text{ critical value} \times SE \text{ of estimate}$$

	parameter	point estimate	SE of estimate	
single proportion	p	$\hat{p}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	when H_0: $p = p_0$, use $\sqrt{\frac{p_0(1-p_0)}{n}}$
diff. of proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	when H_0: $p_1 = p_2$, use $\sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
single mean	μ	$\bar{x}$	$\frac{s}{\sqrt{n}}$	
mean of differences	μ_{diff}	$\bar{x}_{diff}$	$\frac{s_{diff}}{\sqrt{n_{diff}}}$	
difference of means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	
slope of reg. line	β	b	$\frac{s}{s_x\sqrt{n-1}}$	

$$\text{Chi-square test statistic} = \sum\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

CONFIDENCE INTERVALS

Use **confidence intervals** to **estimate** a parameter with a particular **confidence level, C**.

IDENTIFY: Identify the parameter and the confidence level.

CHOOSE: Choose and name the appropriate interval.

CHECK: Check that conditions for the procedure are met.

CALCULATE:
CI: $\textbf{point estimate} \pm \textbf{critical value} \times \textbf{SE of estimate}$

df = (if applicable)
(_____ , _____)

CONCLUDE:
We are C% confident that the true [parameter] is between _____ and _____ . (Put the parameter in *context*.)

We have evidence that [...], because [...]. OR
We do not have evidence that [...], because [...].

HYPOTHESIS TESTS

Use **hypothesis tests** to **test** H_0 versus H_A at a particular **significance level, α**.

IDENTIFY: Identify the hypotheses and the significance level.

CHOOSE: Choose and name the appropriate test.

CHECK: Check that conditions for the procedure are met.

CALCULATE:
$$\textbf{standardized test statistic} = \frac{\textbf{point estimate} - \textbf{null value}}{\textbf{SE of estimate}}$$

df = (if applicable)
p-value =

CONCLUDE:
p-value < α, so we reject H_0.
We have evidence that [H_A]. (Put H_A in *context*.)
        OR
p-value > α, so we do NOT reject H_0.
We do NOT have evidence that [H_A]. (Put H_A in *context*.)

When the parameter is: **a single proportion p**

CHOOSE: **1-Proportion Z-Interval** to estimate p, or
        1-Proportion Z-Test to test $H_0: p = p_0$.

CHECK:
- Data come from a random sample or process.
- for CI: $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$.
  for Test: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

CALCULATE: (`1-PropZInt` or `1-PropZTest`)
point estimate: sample proportion $\hat{p}$

SE of estimate: for CI, use $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; for Test, use $\sqrt{\frac{p_0(1-p_0)}{n}}$

When the parameter is: **a single mean μ**

CHOOSE: **1-Sample T-Interval** to estimate μ, or
        1-Sample T-Test to test $H_0: \mu = \mu_0$.

CHECK:
- Data come from a random sample or process.
- $n \geq 30$, OR population known to be nearly normal, OR population could to be nearly normal because data has no excessive skew or outliers (draw graph).

CALCULATE: (`TInterval` or `T-Test`)
point estimate: sample mean $\bar{x}$
SE of estimate: $\frac{s}{\sqrt{n}}$
$df = n - 1$

When the parameter is: **a difference of proportions p_1-p_2**

CHOOSE: **2-Proportion Z-Interval** to estimate $p_1 - p_2$, or
        2-Proportion Z-Test to test $H_0: p_1 = p_2$.

CHECK:
- Data come from 2 independent random samples or 2 randomly assigned treatments.
- $n_1\hat{p}_1 \geq 10$, $n_1(1 - \hat{p}_1) \geq 10$,
  $n_2\hat{p}_2 \geq 10$, $n_2(1 - \hat{p}_2) \geq 10$.
  Note: use $\hat{p}_c$, the pooled proportion, in place of $\hat{p}_1$ and $\hat{p}_2$ when checking condition for the 2-Proportion Z-Test

CALCULATE: (`2-PropZInt` or `2-PropZTest`)
point estimate: difference of sample proportions $\hat{p}_1 - \hat{p}_2$
SE of estimate:

CI, use $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$;   Test, use $\sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

When the parameter is: **a difference of means μ₁-μ₂**

CHOOSE: **2-Sample T-Interval** to estimate $\mu_1 - \mu_2$, or
        2-Sample T-Test to test $H_0: \mu_1 = \mu_2$.

CHECK:
- Data come from 2 independent random samples or 2 randomly assigned treatments.
- $n_1 \geq 30$ and $n_2 \geq 30$, OR *both* populations known to be nearly normal, OR *both* populations could be nearly normal because both data sets have no excessive skew or outliers (draw 2 graphs).

CALCULATE: (`2-SampTInt` or `2-SampTTest`)
point estimate: difference of sample means $\bar{x}_1 - \bar{x}_2$
SE of estimate: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
df: use technology

CHOOSE: **Matched Pairs T-Interval** to estimate μ_{diff}, or
Matched Pairs T-Test to test $H_0: \mu_{diff} = 0$.

CHECK:
- There is paired data from a random sample or matched pairs experiment.
- $n_{diff} \geq 30$, OR population of differences known to be nearly normal, OR population of differences could be nearly normal because observed differences have no excessive skew or outliers (draw graph of *differences*).

CALCULATE: (`TInterval` or `T-Test`)
point estimate: mean of sample difference $\bar{x}_{diff}$
SE of estimate: $\frac{s_{diff}}{\sqrt{n_{diff}}}$
$df = n_{diff} - 1$

CHOOSE: **T-Interval for the slope** to estimate β, or
T-Test for the slope to test $H_0: \beta = 0$.

CHECK:
- There is (*x*, *y*) data from a random sample or experiment.
- The residual plot shows no pattern making a linear model reasonable. (More specifically, the residuals should be independent, nearly normal, and have constant standard deviation.)

CALCULATE: (`LinRegTInt` or `LinRegTTest`)
point estimate: sample slope b
SE of estimate: SE of slope (from computer output)
$df = n - 2$

The χ2 tests for categorical variables: chi-square statistic $= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

CHOOSE: **χ2 Goodness of Fit Test**

CHECK:
- Data come from a random sample or process.
- All expected counts ≥ 5. (To calculate expected counts for each category, multiply the sample size by the expected proportion under H_0.)

CALCULATE: (`χ2GOF-Test`)
$\chi^2 =$
$df = \text{\# of categories} - 1$

CHOOSE: **χ2 Test for Homogeneity**

CHECK:
- Data come from 2 or more independent random samples or 2 or more randomly assigned treatments.
- All expected counts ≥ 5. (Calculate expected counts and verify this to be true.)

CALCULATE: (`χ2-Test`, then `2ND MATRIX, EDIT, 2:[B]` to find expected counts)
$\chi^2 =$
$df = (\text{\# of rows} - 1)(\text{\# of cols} - 1)$

CHOOSE: **χ2 Test for Independence**

CHECK:
- Data come from a random sample or process.
- All expected counts ≥ 5. (Calculate expected counts and verify this to be true.)

CALCULATE: (`χ2-Test`, then `2ND MATRIX, EDIT, 2:[B]` to find expected counts)
$\chi^2 =$
$df = (\text{\# of rows} - 1)(\text{\# of cols} - 1)$