Introductory Statistics with Randomization and Simulation

First Edition

David M Diez
Quantitative Analyst
Google/YouTube
david@openintro.org

Christopher D Barr
Graduate Student
Yale School of Management
chris@openintro.org

Mine Çetinkaya-Rundel
Assistant Professor of the Practice
Department of Statistics
Duke University
mine@openintro.org

Contents

Preface

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

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 interesting, real data.

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

Textbook overview

The chapters of this book are as follows:

1. Introduction to data. Data structures, variables, summaries, graphics, and basic data collection techniques.

2. Foundations for inference. Case studies are used to introduce the ideas of statistical inference with randomization and simulations. The content leads into the standard parametric framework, with techniques reinforced in the subsequent chapters.[1] It is also possible to begin with this chapter and introduce tools from Chapter 1 as they are needed.

3. Inference for categorical data. Inference for proportions using the normal and chi-square distributions, as well as simulation and randomization techniques.

4. Inference for numerical data. Inference for one or two sample means using the t distribution, and also comparisons of many means using ANOVA. A special section for bootstrapping is provided at the end of the chapter.

5. Introduction to linear regression. An introduction to regression with two variables. Most of this chapter could be covered immediately after Chapter 1.

6. Multiple and logistic regression. An introduction to multiple regression and logistic regression for an accelerated course.

Appendix A. Probability. An introduction to probability is provided as an optional reference. Exercises and additional probability content may be found in Chapter 2 of *OpenIntro Statistics* at **openintro.org**. Instructor feedback suggests that probability, if discussed, is best introduced at the very start or very end of the course.

[1] Instructors who have used similar approaches in the past may notice the absence of the bootstrap. Our investigation of the bootstrap has shown that there are many misunderstandings about its robustness. For this reason, we postpone the introduction of this technique.

Examples, exercises, and additional appendices

Examples and guided practice exercises throughout the textbook may be identified by their distinctive bullets:

● **Example 0.1** Large filled bullets signal the start of an example.

Full solutions to examples are provided and often include an accompanying table or figure.

⊙ **Guided Practice 0.2** Large empty bullets signal to readers that an exercise has been inserted into the text for additional practice and guidance. Solutions for all guided practice exercises are provided in footnotes.[2]

Exercises at the end of each chapter are useful for practice or homework assignments. Many of these questions have multiple parts, and solutions to odd-numbered exercises can be found in Appendix B.

Probability tables for the normal, t, and chi-square distributions are in Appendix C, and PDF copies of these tables are also available from **openintro.org**.

OpenIntro, online resources, and getting involved

OpenIntro is an organization focused on developing free and affordable education materials. We encourage anyone learning or teaching statistics to visit **openintro.org** and get involved. We also provide many free online resources, including free course software. Data sets for this textbook are available on the website and through a companion R package.[3] All of these resources are free, and we want to be clear that anyone is welcome to use these online tools and resources with or without this textbook as a companion.

We value your feedback. If there is a part of the project you especially like or think needs improvement, we want to hear from you. You may find our contact information on the title page of this book or on the About section of **openintro.org**.

Acknowledgements

This project would not be possible without the dedication and volunteer hours of all those involved. No one has received any monetary compensation from this project, and we hope you will join us in extending a *thank you* to all those who volunteer with OpenIntro.

The authors would especially like to thank Andrew Bray and Meenal Patel for their involvement and contributions to this textbook. We are also grateful to Andrew Bray, Ben Baumer, and David Laffie for providing us with valuable feedback based on their experiences while teaching with this textbook, and to the many teachers, students, and other readers who have helped improve OpenIntro resources through their feedback.

The authors would like to specially thank George Cobb of Mount Holyoke College and Chris Malone of Winona State University. George has spent a good part of his career supporting the use of nonparametric techniques in introductory statistics, and Chris was helpful in discussing practical considerations for the ordering of inference used in this textbook. Thank you, George and Chris!

[2] Full solutions are located down here in the footnote!

[3] Diez DM, Barr CD, Çetinkaya-Rundel M. 2012. `openintro`: OpenIntro data sets and supplement functions. http://cran.r-project.org/web/packages/openintro.

Chapter 1

Introduction to data

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.

Statistics as a subject focuses on making stages 2-4 objective, rigorous, and efficient. That is, statistics has three primary components: How best can we collect data? How should it be analyzed? And what can we infer from the analysis?

The topics scientists investigate are as diverse as the questions they ask. However, many of these investigations can be addressed with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference. This chapter provides a glimpse into these and other themes we will encounter throughout the rest of the book. We introduce the basic principles of each branch and learn some tools along the way. We will encounter applications from other fields, some of which are not typically associated with science but nonetheless can benefit from statistical study.

1.1 Case study: using stents to prevent strokes

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 small mesh tubes that are placed inside

[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. http://www.nejm.org/doi/full/10.1056/NEJMoa1105335. NY Times article reporting on the study: http://www.nytimes.com/2011/09/08/health/research/08stent.html.

narrow or weak arteries to 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 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.

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 Table 1.1. Patient outcomes are recorded as "stroke" or "no event".

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

Table 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, a statistical analysis allows us to consider all of the data at once. Table 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

Table 1.2: Descriptive statistics for the stent study.

⊙ **Guided Practice 1.1** Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Answers to all in-text exercises are provided in 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 due to the treatment?

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 the following: is the difference so large that we should reject the notion that it was due to chance?

While we haven't yet covered statistical tools to fully address this question, 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.

1.2 Data basics

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.

1.2.1 Observations, variables, and data matrices

Table 1.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 emails received in 2012. These observations will be referred to as the `email50` data set, and they are a random sample from a larger data set that we will see in Section 1.7.

[2] The proportion of the 224 patients who had a stroke within 365 days: $45/224 = 0.20$.

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

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 1.3: Four rows from the `email50` data matrix.

variable	description
`spam`	Specifies whether the message was spam
`num_char`	The number of characters in the email
`line_breaks`	The number of line breaks in the email (not including text wrapping)
`format`	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
`number`	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Table 1.4: Variables and their descriptions for the `email50` data set.

Each row in the table represents a single email or **case**.[4] The columns represent characteristics, called **variables**, for each of the emails. For example, the first row represents email 1, which is not spam, contains 21,705 characters, 551 line breaks, is written in HTML format, and contains only small numbers.

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 all five email variables are given in Table 1.4.

The data in Table 1.3 represent a **data matrix**, which is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A data matrix for the stroke study introduced in Section 1.1 is shown in Table 1.1 on page 2, where the cases were patients and there were three variables recorded for each patient.

Data matrices are a convenient way to record and store data. If another individual or case is added to the data set, an additional row can be easily added. Similarly, another column can be added for a new variable.

⊙ **Guided Practice 1.2** We consider a publicly available data set that summarizes information about the 3,143 counties in the United States, and we call this the `county` data set. This data set includes information about each county: its name, the state where it resides, its population in 2000 and 2010, per capita federal spending, poverty rate, and five additional characteristics. How might these data be organized in a data matrix? Reminder: look in the footnotes for answers to in-text exercises.[5]

Seven rows of the `county` data set are shown in Table 1.5, and the variables are summarized in Table 1.6. These data were collected from the US Census website.[6]

[4] A case is also sometimes called a **unit of observation** or an **observational unit**.

[5] Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,143 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

[6] http://quickfacts.census.gov/qfd/index.html

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

Table 1.5: Seven rows from the `county` data set.

variable	description
`name`	County name
`state`	State where the county resides (also including the District of Columbia)
`pop2000`	Population in 2000
`pop2010`	Population in 2010
`fed_spend`	Federal spending per capita
`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)
`multiunit`	Percent of living units that are in multi-unit structures (e.g. apartments)
`income`	Income per capita
`med_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
`smoking_ban`	Type of county-wide smoking ban in place at the end of 2011, which takes one of three values: `none`, `partial`, or `comprehensive`, where a `comprehensive` ban means smoking was not permitted in restaurants, bars, or workplaces, and `partial` means smoking was banned in at least one of those three locations

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

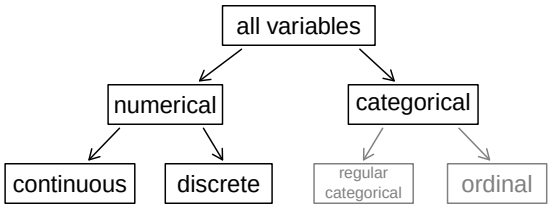


Figure 1.7: Breakdown of variables into their respective types.

1.2.2 Types of variables

Examine the `fed_spend`, `pop2010`, `state`, and `smoking_ban` variables in the `county` data set. Each of these variables is inherently different from the other three yet many of them share certain characteristics.

First consider `fed_spend`, 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 their average, sum, and difference have no clear meaning.

The `pop2010` variable is also numerical, although it seems to be a little different than `fed_spend`. This variable of the population count can only be a whole non-negative number (`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 federal spending variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: `AL`, ..., and `WY`. Because the responses themselves are categories, `state` is called a **categorical** variable,[7] and the possible values are called the variable's **levels**.

Finally, consider the `smoking_ban` variable, which describes the type of county-wide smoking ban and takes a value `none`, `partial`, or `comprehensive` 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. To simplify analyses, any ordinal variables in this book will be treated as categorical variables.

Example 1.3 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 1.4 Consider the variables `group` and `outcome` (at 30 days) from the stent study in Section 1.1. Are these numerical or categorical variables?[8]

[7] Sometimes also called a **nominal** variable.

[8] There are only two possible values for each variable, and in both cases they describe categories. Thus, each is a categorical variable.

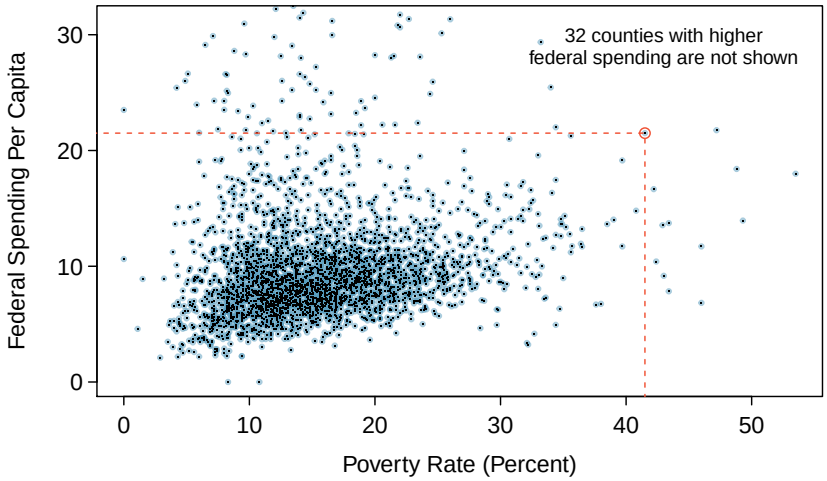


Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

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) Is federal spending, on average, higher or lower in counties with high rates of poverty?

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

(3) Which counties have a higher average income: those that enact one or more smoking bans or those that do not?

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

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables `fed_spend` and `poverty`. Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 1088 in the `county` data set: Owsley County, Kentucky, which had a poverty rate of 41.5% and federal spending of $21.50 per capita. The dense cloud in the scatterplot suggests a relationship between the two variables: counties with a high poverty rate also tend to have slightly more federal spending. We might brainstorm as to why this relationship exists and investigate each idea to determine which is the most reasonable explanation.

⊙ **Guided Practice 1.5** Examine the variables in the `email50` data set, which are described in Table 1.4 on page 4. Create two questions about the relationships between these variables that are of interest to you.[9]

[9] Two sample questions: (1) Intuition suggests that if there are many line breaks in an email then there would also tend to be many characters: does this hold true? (2) Is there a connection between whether an email format is plain text (versus HTML) and whether it is a spam message?

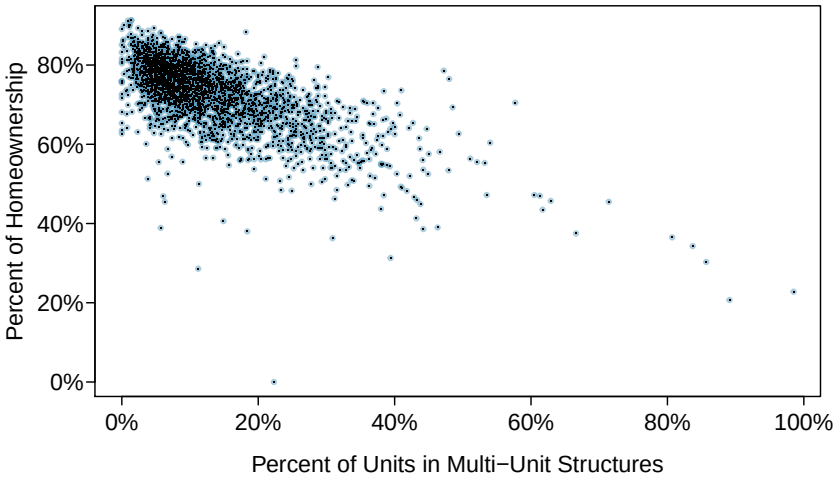


Figure 1.9: A scatterplot of the homeownership rate versus the percent of units that are in multi-unit structures for all 3,143 counties. Interested readers may find an image of this plot with an additional third variable, county population, presented at www.openintro.org/stat/down/MHP.png.

The `fed_spend` and `poverty` variables 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.

Example 1.6 The relationship between the homeownership rate and the percent of units in multi-unit structures (e.g. apartments, condos) is visualized using a scatterplot in Figure 1.9. Are these variables associated?

It appears that the larger the fraction of units in multi-unit structures, the lower the homeownership rate. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.9 – 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 `poverty` and `fed_spend` variables represented in Figure 1.8, where counties with higher poverty rates tend to receive more federal spending per capita.

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

1.3 Overview of data collection principles

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

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 undergraduate students?
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. It is usually 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.

⊙ **Guided Practice 1.7** For the second and third questions above, identify the target population and what represents an individual case.[10]

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**.

[10] (2) Notice that the first 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 undergraduate students who have graduated in the last five years represent cases in the population under consideration. 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.

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

February 10th, 2010.

Anecdotal evidence
Be careful of data collected haphazardly. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

1.3.3 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*. In general, we always seek to *randomly* select a sample from a population. 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.

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

Example 1.8 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 collect? 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.

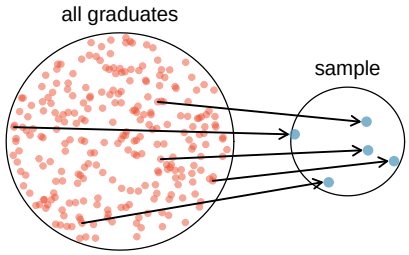


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

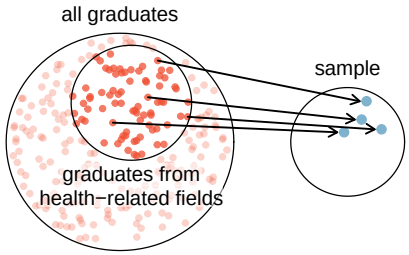


Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample. Sampling randomly helps resolve this problem. 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.

The act of taking a simple 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, and it is unclear whether the respondents are **representative** of the entire population, the survey might suffer from **non-response bias**.

Another 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, it will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

⊙ **Guided Practice 1.9** 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?[11]

[11] 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.

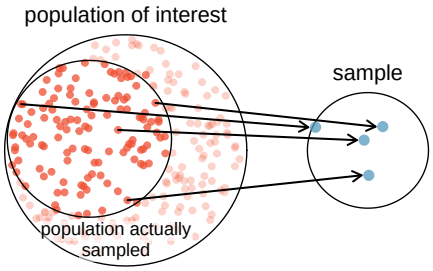


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often impossible, to completely fix this problem.

1.3.4 Explanatory and response variables

Consider the following question from page 7 for the `county` data set:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

If we suspect poverty might affect spending in a county, then poverty is the **explanatory** variable and federal spending is the **response** variable in the relationship.[12] If there are many variables, it may be possible to consider a number of them as explanatory variables.

TIP: Explanatory and response variables
To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other.

explanatory variable → might affect → response variable

Caution: 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 some cases, there is no explanatory or response variable. Consider the following question from page 7:

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

It is difficult to decide which of these variables should be considered the explanatory and response variable, i.e. the direction is ambiguous, so no explanatory or response labels are suggested here.

[12] 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.

1.3.5 Introducing observational studies and experiments

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

Researchers perform an **observational study** when they collect data in a way that does not directly interfere 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 what happens. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. Usually 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, perhaps by flipping a coin, 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.

TIP: association $\neq$ causation
In a data analysis, association does not imply causation, and causation can only be inferred from a randomized experiment.

1.4 Observational studies and sampling strategies

1.4.1 Observational studies

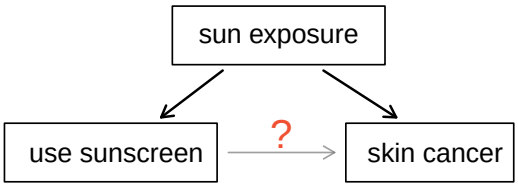
Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.

Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations.

⊙ **Guided Practice 1.10** Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[13]

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. If someone is out in the sun all day, she is more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.

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



Sun exposure is what is called a **confounding variable**,[14] which is a variable that is correlated with both the explanatory and response variables. 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 easily be used to make causal conclusions.

⊙ **Guided Practice 1.11** Figure 1.9 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 in Figure 1.9. [15]

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.[16] 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).

1.4.2 Three sampling methods (special topic)

Almost all statistical methods are based on the notion of implied randomness. If observational data are not collected in a random framework from a population, results from these statistical methods are not reliable. Here we consider three random sampling techniques: simple, stratified, and cluster sampling. Figure 1.14 provides a graphical representation of these techniques.

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. To take a simple random sample of 120 baseball players and their salaries from the 2010 season, we could write the names of that season's

[14] Also called a **lurking variable**, **confounding factor**, or a **confounder**.

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

[16] http://www.channing.harvard.edu/nhs/

Figure 1.14: Examples of simple random, stratified, and cluster sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the middle panel, stratified sampling was used: cases were grouped into strata, and then simple random sampling was employed within each stratum. In the bottom panel, cluster sampling was used, where data were binned into nine clusters, and three of the clusters were randomly selected.

828 players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as "simple random" if each case in the population has an equal chance of being included in the final sample *and* knowing that a case is included in a sample does not provide useful information about which other cases are included.

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, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata; some teams have a lot more money (we're looking at you, Yankees). Then we might randomly sample 4 players from each team for a total of 120 players.

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample. The analysis methods introduced in this book would need to be extended to analyze data collected using stratified sampling.

Example 1.12 Why would it be good for cases within each stratum to be very similar?

We might 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.

In **cluster sampling**, we group observations into clusters, then randomly sample some of the clusters. Sometimes cluster sampling can be a more economical technique than the alternatives. Also, 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. For example, if neighborhoods represented clusters, then this sampling method works best when the neighborhoods are very diverse. A downside of cluster sampling is that more advanced analysis techniques are typically required, though the methods in this book can be extended to handle such data.

Example 1.13 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. 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, cluster sampling seems like a very good idea. We might randomly select a small number of villages. This would probably reduce our data collection costs substantially in comparison to a simple random sample and would still give us helpful information.

Another technique called **multistage sampling** is similar to cluster sampling, except that we take a simple random sample within each selected cluster. For instance, if we sampled neighborhoods using cluster sampling, we would next sample a subset of homes within each selected neighborhood if we were using multistage sampling.

1.5 Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

1.5.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. 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.

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 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 a single study, we **replicate** by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients into low-risk and high-risk **blocks**, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.15. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate the first three experimental design principles into any study, and this book describes methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in this book may be extended to analyze data collected using blocking.

1.5.2 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationships 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.[17] In particular, researchers wanted to know if the drug reduced deaths in patients.

[17] 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.

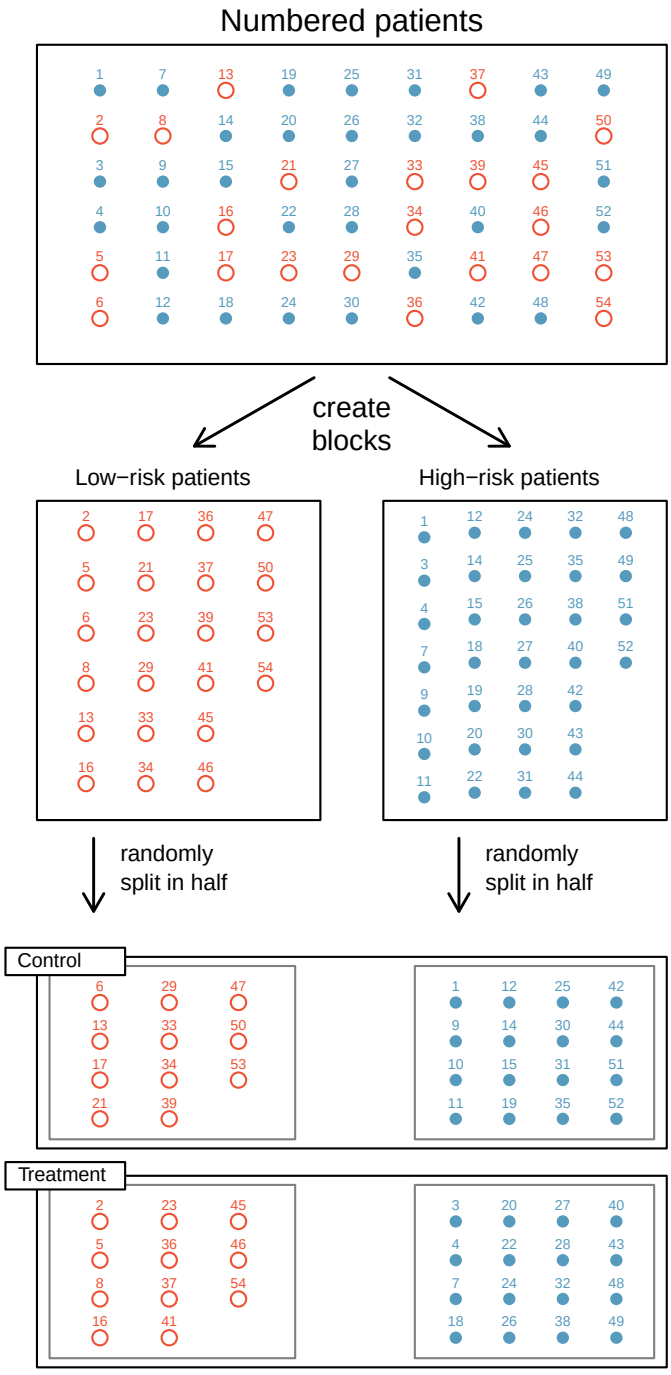


Figure 1.15: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly divided into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[18] 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.

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**. 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 doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[19]

⊙ **Guided Practice 1.14** 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?[20]

1.6 Examining numerical data

This section introduces techniques for exploring and summarizing numerical variables, and the `email50` and `county` data sets from Section 1.2 provide rich opportunities for examples. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations. For example, the `pop2010` variable, which represents the populations of counties in 2010, is numerical since we can sensibly discuss the difference or ratio of the populations in two counties. On the other hand, area codes and zip codes are not numerical.

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

[19] There are always some researchers 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.

[20] 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.

1.6.1 Scatterplots for paired data

A **scatterplot** provides a case-by-case view of data for two numerical variables. In Figure 1.8 on page 7, a scatterplot was used to examine how federal spending and poverty were related in the `county` data set. Another scatterplot is shown in Figure 1.16, comparing 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 1.16.

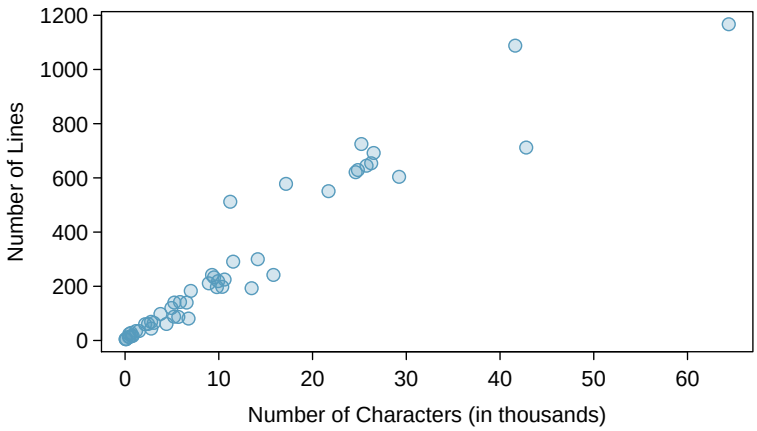


Figure 1.16: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

To put the number of characters in perspective, this paragraph has 363 characters. Looking at Figure 1.16, it seems that some emails are incredibly long! Upon further investigation, we would actually find that most of the long emails use the HTML format, which means most of the characters in those emails are used to format the email rather than provide text.

⊙ **Guided Practice 1.15** What do scatterplots reveal about the data, and how might they be useful?[21]

● **Example 1.16** Consider a new data set of 54 cars with two variables: vehicle price and weight.[22] A scatterplot of vehicle price versus weight is shown in Figure 1.17. 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, such as Figure 1.8 on page 7 and Figure 1.16, which show relationships that are very linear.

⊙ **Guided Practice 1.17** Describe two variables that would have a horseshoe shaped association in a scatterplot.[23]

[21] Answers may vary. Scatterplots are helpful in quickly spotting associations between variables, whether those associations represent simple or more complex relationships.

[22] Subset of data from http://www.amstat.org/publications/jse/v1n1/datasets.lock.html

[23] 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 since water becomes toxic when consumed in excessive quantities.

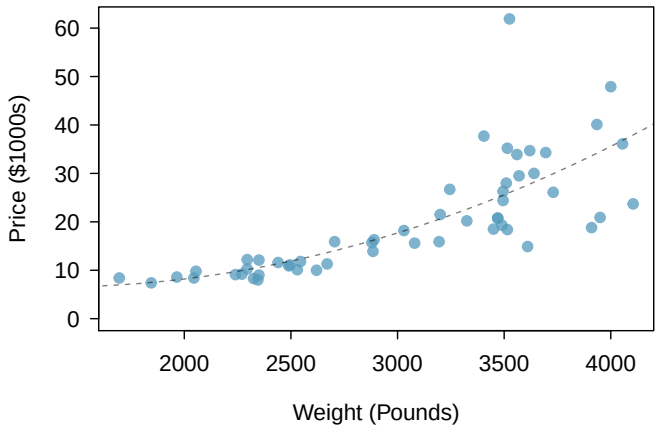


Figure 1.17: A scatterplot of `price` versus `weight` for 54 cars.

1.6.2 Dot plots and the mean

Sometimes two variables is one too many: only one variable may be of interest. In these cases, a dot plot provides the most basic of displays. A **dot plot** is a one-variable scatterplot; an example using the number of characters from 50 emails is shown in Figure 1.18. A stacked version of this dot plot is shown in Figure 1.19.

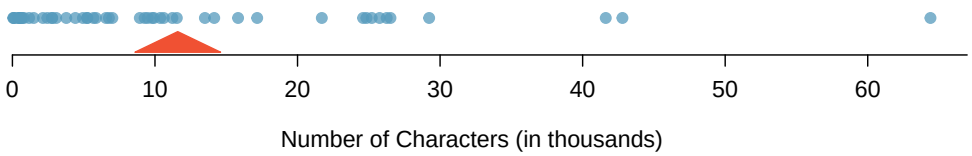


Figure 1.18: A dot plot of `num_char` for the `email50` data set.

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.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \tag{1.18}$$

The sample mean is often labeled $\bar{x}$, and the letter x is being used as a generic placeholder for the variable of interest, `num_char`. The sample mean is shown as a triangle in Figures 1.18 and 1.19.

$\bar{x}$ sample mean

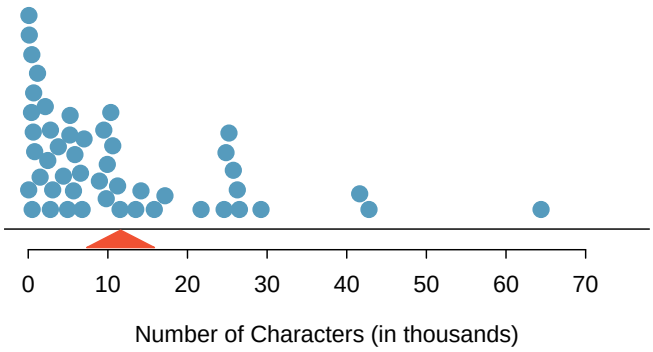


Figure 1.19: A stacked dot plot of `num_char` for the `email50` data set.

Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.19}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

n
sample size

⊙ **Guided Practice 1.20** Examine Equations (1.18) and (1.19) above. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[24]

⊙ **Guided Practice 1.21** What was n in this sample of emails?[25]

The `email50` data set is 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, there is a difference in notation: 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.

μ
population mean

● **Example 1.22** 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 2 and beyond, we will develop tools to characterize the accuracy of point estimates, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

[24] 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 i^{th} email in the data set.

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

Example 1.23 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 from the 3,143 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 1.23 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:

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

1.6.3 Histograms and shape

Dot plots show the exact value of each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, 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. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower bin. This tabulation is shown in Table 1.20. These binned counts are plotted as bars in Figure 1.21 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 1.19.

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

Table 1.20: The counts for the binned `num_char` data.

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

Histograms are especially convenient for describing the shape of the data distribution. Figure 1.21 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**.[26]

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**.

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

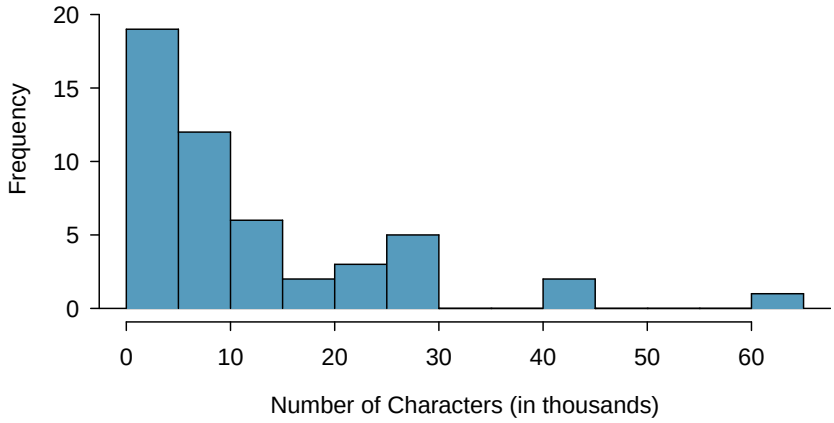


Figure 1.21: A histogram of `num_char`. This distribution is very strongly skewed to the right.

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 1.24** Take a look at the dot plots in Figures 1.18 and 1.19. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[27]

⊙ **Guided Practice 1.25** Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[28]

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes. A **mode** is represented by a prominent peak in the distribution.[29] There is only one prominent peak in the histogram of `num_char`.

Figure 1.22 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 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.

⊙ **Guided Practice 1.26** Figure 1.21 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[30]

[27] The skew is visible in all three plots, though the flat dot plot is the least useful. The stacked dot plot and histogram are helpful visualizations for identifying skew.

[28] Character counts for individual emails.

[29] 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.

[30] Unimodal. Remember that *uni* stands for 1 (think *uni*cycles). Similarly, *bi* stands for 2 (think *bi*cycles). (We're hoping a *multicycle* will be invented to complete this analogy.)

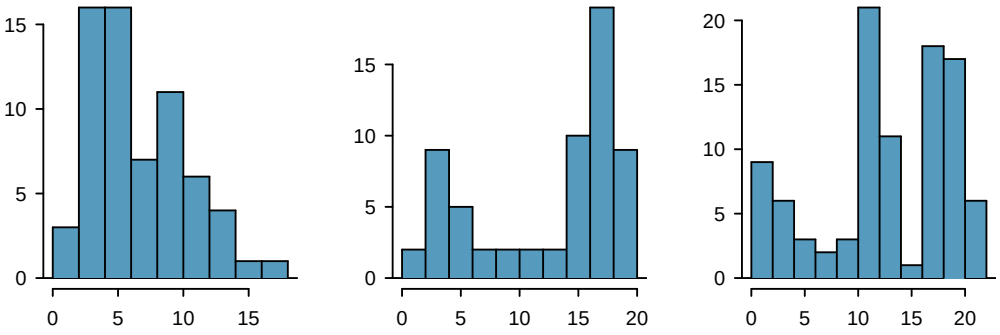


Figure 1.22: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

⊙ **Guided Practice 1.27** 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?[31]

TIP: 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.

1.6.4 Variance and standard deviation

The mean is used to describe the center of a data set, but the variability in the data is also important. Here, we introduce two measures of variability: the variance and the standard deviation. Both of these are very useful in data analysis, even though the formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to conceptually understand, and it roughly describes how far away the typical observation is from the mean.

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

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

[31] 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.

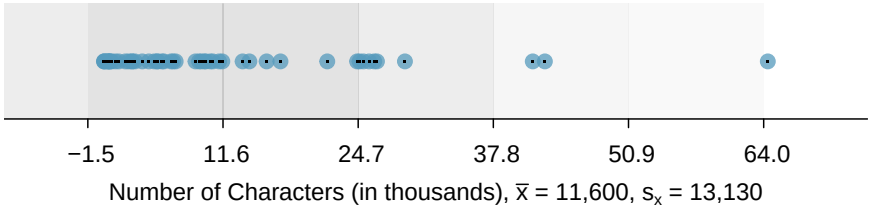


Figure 1.23: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

s^2
sample variance

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

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

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 the square root of the variance:

s
sample standard deviation

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

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.

Variance and standard deviation
The variance is roughly the average squared distance from the mean. The standard deviation is the square root of the variance and describes how close the data are to the mean.

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[32] 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*.

σ^2
population variance

σ
population standard deviation

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

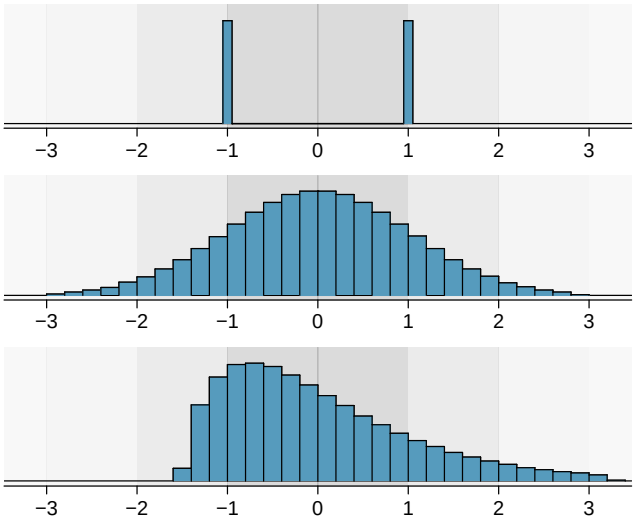


Figure 1.24: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

TIP: standard deviation describes variability
Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 1.23 and 1.24, these percentages are not strict rules.

Guided Practice 1.28 On page 23, 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 1.24 as an example, explain why such a description is important.[33]

Example 1.29 Describe the distribution of the `num_char` variable using the histogram in Figure 1.21 on page 24. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.

The distribution of email character counts is unimodal and very strongly skewed to the high end. 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 practice, the variance and standard deviation are sometimes used as a means to an end, where the "end" is being able to accurately estimate the uncertainty associated with a sample statistic. For example, in Chapter 2 we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

[33] Figure 1.24 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.

1.6.5 Box plots, quartiles, and the median

A **box plot** summarizes a data set using five statistics while also plotting unusual observations. Figure 1.25 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email50` data set.

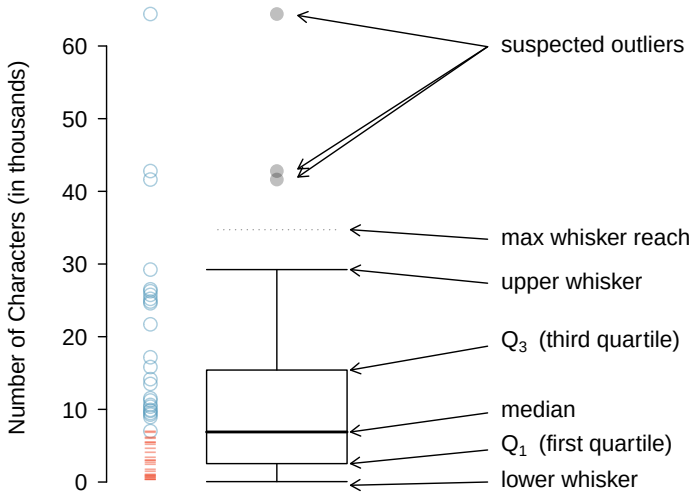


Figure 1.25: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 1.25 shows 50% of the data falling below the median (dashes) and other 50% falling above the median (open circles). There are 50 character counts in the 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 observations closest to the 50^{th} percentile: $(6{,}768 + 7{,}012)/2 = 6{,}890$. 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
If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 1.25, is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR. The two boundaries of the box are called the **first quartile** (the 25^{th} percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75^{th} percentile), and these are often labeled Q_1 and Q_3, respectively.

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 25^{th} and 75^{th} percentiles.

⊙ **Guided Practice 1.30** What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[34]

Extending out from the box, the **whiskers** attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$.[35] They capture everything within this reach. In Figure 1.25, the upper whisker does not extend to the last three points, which are beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 33, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. 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.

Any observation that lies beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of just extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called **outliers**. In this case, it would be reasonable to classify the emails with character counts of 41,623, 42,793, and 64,401 as outliers since they are numerically distant from most of the data.

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

TIP: Why it is important to look for outliers
Examination of data for possible outliers serves many useful purposes, including

1. Identifying strong skew in the distribution.
2. Identifying data collection or entry errors. For instance, we re-examined the email purported to have 64,401 characters to ensure this value was accurate.
3. Providing insight into interesting properties of the data.

⊙ **Guided Practice 1.31** The observation 64,401, an outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[36]

⊙ **Guided Practice 1.32** Using Figure 1.25, estimate the following values for `num_char` in the `email50` data set: (a) Q_1, (b) Q_3, and (c) IQR.[37]

[34] Since Q_1 and Q_3 capture the middle 50% of the data and the median splits the data in the middle, 25% of the data fall between Q_1 and the median, and another 25% falls between the median and Q_3.

[35] While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.

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

[37] These visual estimates will vary a little from one person to the next: $Q_1 \approx 3{,}000$, $Q_3 \approx 15{,}000$, IQR $= Q_3 - Q_1 \approx 12{,}000$. (The true values: $Q_1 = 2{,}536$, $Q_3 = 15{,}411$, IQR $= 12{,}875$.)

1.6.6 Robust statistics

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 1.26, and sample statistics are computed under each scenario in Table 1.27.

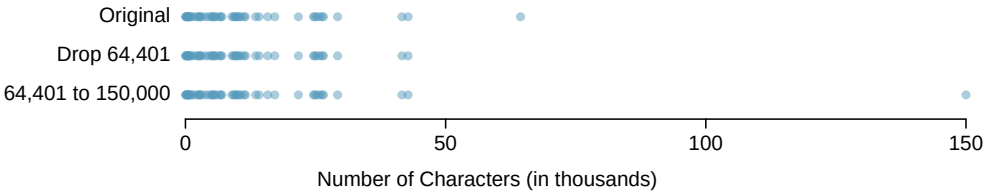


Figure 1.26: Dot plots of the original character count data and two modified data sets.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `num_char` data	6,890	12,875	11,600	13,130
drop 66,924 observation	6,768	11,702	10,521	10,798
move 66,924 to 150,000	6,890	12,875	13,310	22,434

Table 1.27: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

Guided Practice 1.33 (a) Which is more affected by extreme observations, the mean or median? Table 1.27 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[38]

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 1.34 The median and IQR do not change much under the three scenarios in Table 1.27. Why might this be the case?

The median and IQR are only sensitive to numbers near Q_1, the median, and Q_3. Since values in these regions are relatively stable – there aren't large jumps between observations – the median and IQR estimates are also quite stable.

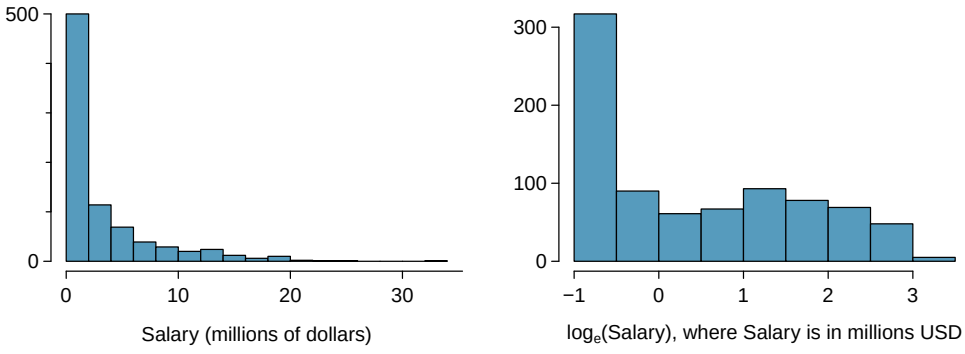
Guided Practice 1.35 The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[39]

[38] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 1.33.

[39] 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.

1.6.7 Transforming data (special topic)

When data are very strongly skewed, we sometimes transform them so they are easier to model. Consider the histogram of salaries for Major League Baseball players' salaries from 2010, which is shown in Figure 1.28(a).



(a) (b)

Figure 1.28: (a) Histogram of MLB player salaries for 2010, in millions of dollars. (b) Histogram of the log-transformed MLB player salaries for 2010.

Example 1.36 The histogram of MLB player salaries is useful in that we can see the data are extremely skewed and centered (as gauged by the median) at about \$1 million. What isn't useful about this plot?

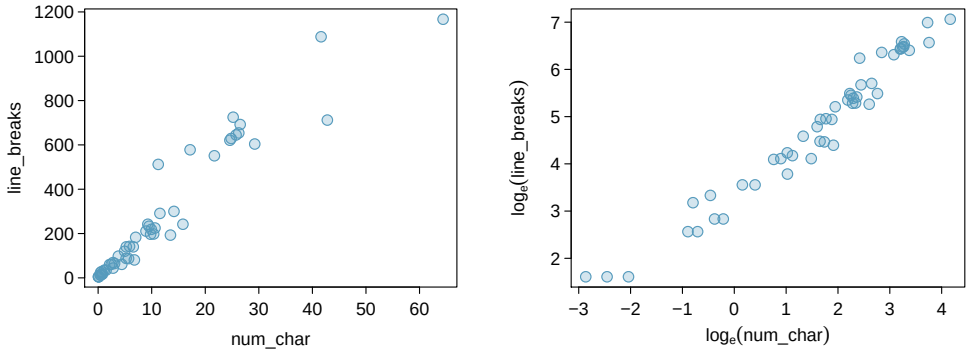
Most of the data are collected into one bin in the histogram and the data are so strongly skewed that many details in the data are obscured.

There are some standard transformations that are often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive. A **transformation** is a rescaling of the data using a function. For instance, a plot of the natural logarithm[40] of player salaries results in a new histogram in Figure 1.28(b). Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the `line_breaks` and `num_char` variables is shown in Figure 1.29(a), which was earlier shown in Figure 1.16. We can see a positive association between the variables and that many observations are clustered near zero. In Chapter 5, we might want to use a straight line to model the data. However, we'll find that the data in their current state cannot be modeled very well. Figure 1.29(b) shows a scatterplot where both the `line_breaks` and `num_char` variables have been transformed using a log (base e) transformation. While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.

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 used by statisticians. 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.

[40] Statisticians often write the natural logarithm as log. You might be more familiar with it being written as ln.



(a) (b)

Figure 1.29: (a) Scatterplot of `line_breaks` against `num_char` for 50 emails. (b) A scatterplot of the same data but where each variable has been log-transformed.

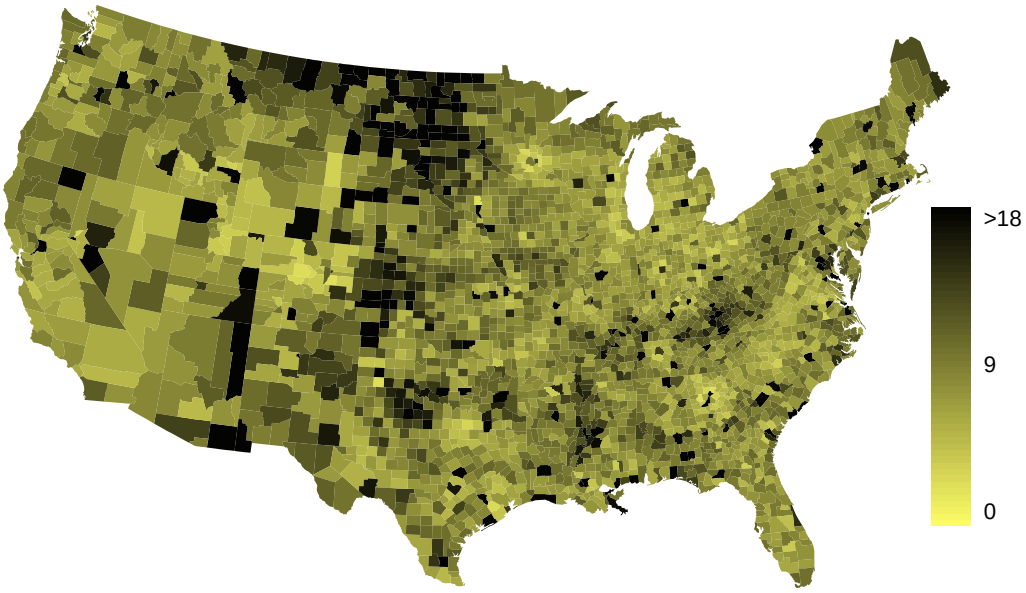
1.6.8 Mapping data (special topic)

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 map it using an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 1.30 and 1.31 shows intensity maps for federal spending per capita (`fed_spend`), poverty rate in percent (`poverty`), homeownership rate in percent (`homeownership`), and median household income (`med_income`). The color key indicates which colors correspond to which values. Note that 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.

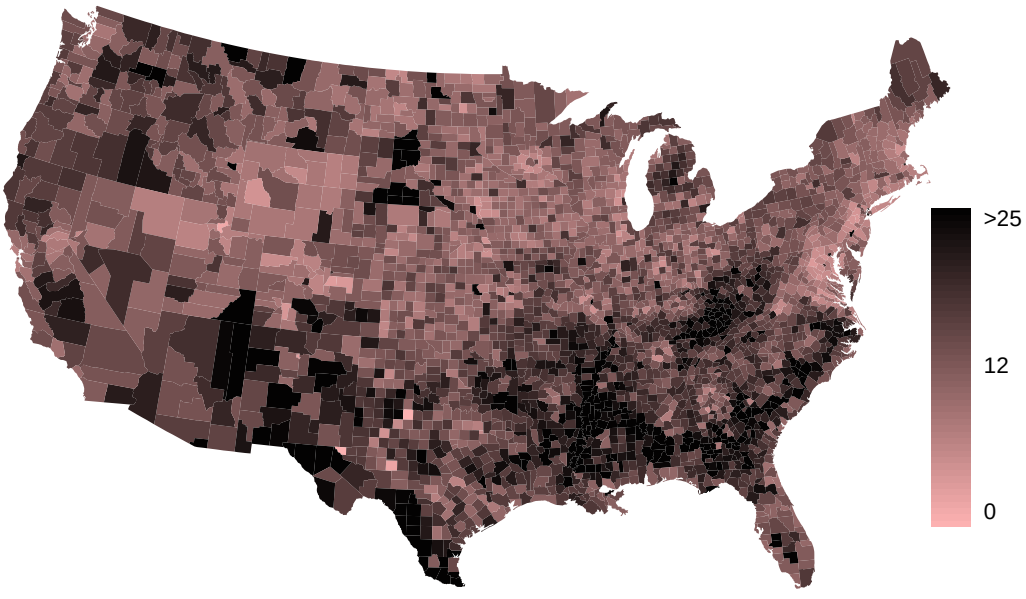
Example 1.37 What interesting features are evident in the `fed_spend` and `poverty` intensity maps?

The federal spending intensity map shows substantial spending in the Dakotas and along the central-to-western part of the Canadian border, which may be related to the oil boom in this region. There are several other patches of federal spending, such as a vertical strip in eastern Utah and Arizona and the area where Colorado, Nebraska, and Kansas meet. There are also seemingly random counties with very high federal spending relative to their neighbors. If we did not cap the federal spending range at $18 per capita, we would actually find that some counties have extremely high federal spending while there is almost no federal spending in the neighboring counties. These high-spending counties might contain military bases, companies with large government contracts, or other government facilities with many employees.

Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does the southwest border of Texas. The vertical strip of eastern Utah and Arizona, noted above for its higher federal spending, also appears to have higher rates of poverty (though generally little correspondence is seen between the two variables). High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky and West Virginia.

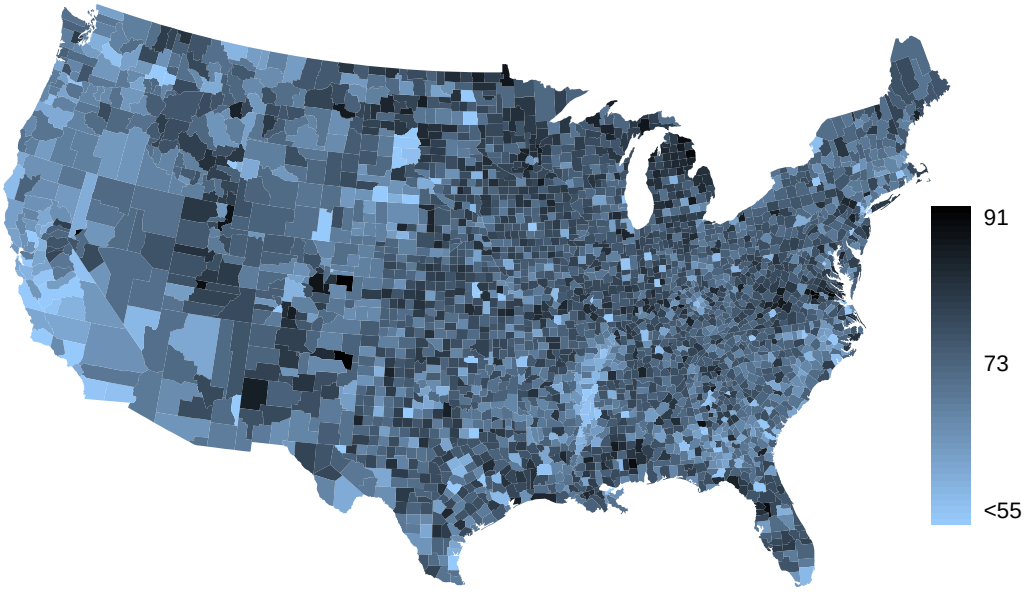


(a)

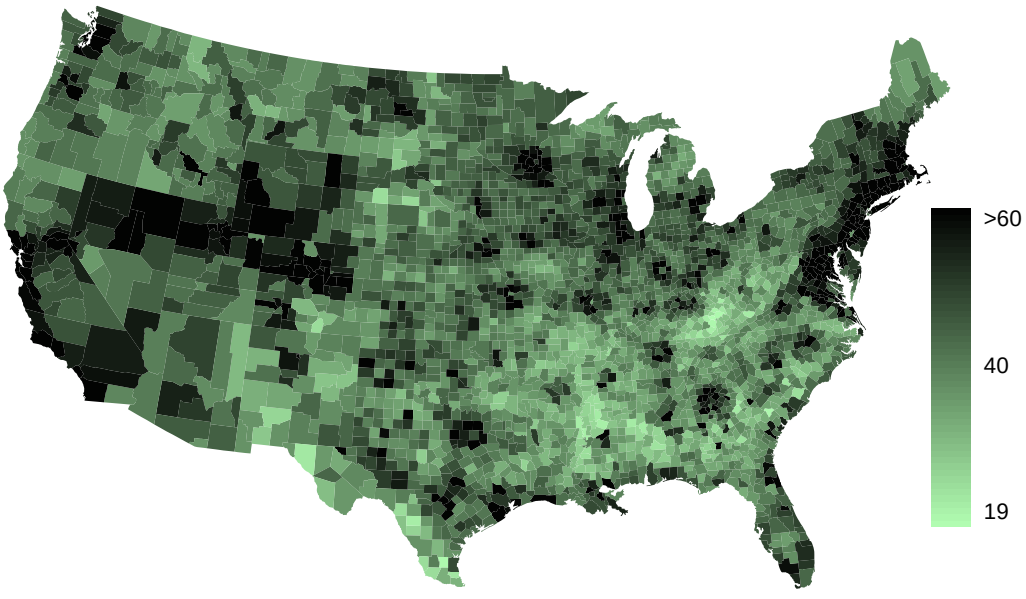


(b)

Figure 1.30: (a) Map of federal spending (dollars per capita). (b) Intensity map of poverty rate (percent).



(a)



(b)

Figure 1.31: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income ($1000s).

⊙ **Guided Practice 1.38** What interesting features are evident in the `med_income` intensity map?[41]

1.7 Considering categorical data

Like numerical data, categorical data can also be organized and analyzed. This section introduces tables and other basic tools for categorical data that are used throughout this book. The `email50` data set represents a sample from a larger email data set called `email`. This larger data set contains information on 3,921 emails. In this section we will examine whether the presence of numbers, small or large, in an email provides any useful value in classifying email as spam or not spam.

1.7.1 Contingency tables and bar plots

Table 1.32 summarizes two variables: `spam` and `number`. Recall that `number` is a categorical variable that describes whether an email contains no numbers, only small numbers (values under 1 million), or at least one big number (a value of 1 million or more). 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 149 corresponds to the number of emails in the data set that are spam *and* had no number listed in the email. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $149 + 168 + 50 = 367$), and **column totals** are total counts down each column.

A table for a single variable is called a **frequency table**. Table 1.33 is a frequency table for the `number` variable. If we replaced the counts with percentages or proportions, the table would be called a **relative frequency table**.

		`number`			
		none	small	big	Total
`spam`	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

Table 1.32: A contingency table for `spam` and `number`.

none	small	big	Total
549	2827	545	3921

Table 1.33: A frequency table for the `number` variable.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 1.34 shows a **bar plot** for the `number` variable. In the right panel, the counts are converted into proportions (e.g. $549/3921 = 0.140$ for `none`).

[41] Note: answers will vary. There is a very strong correspondence between high earning and metropolitan areas. You might look for large cities you are familiar with and try to spot them on the map as dark spots.

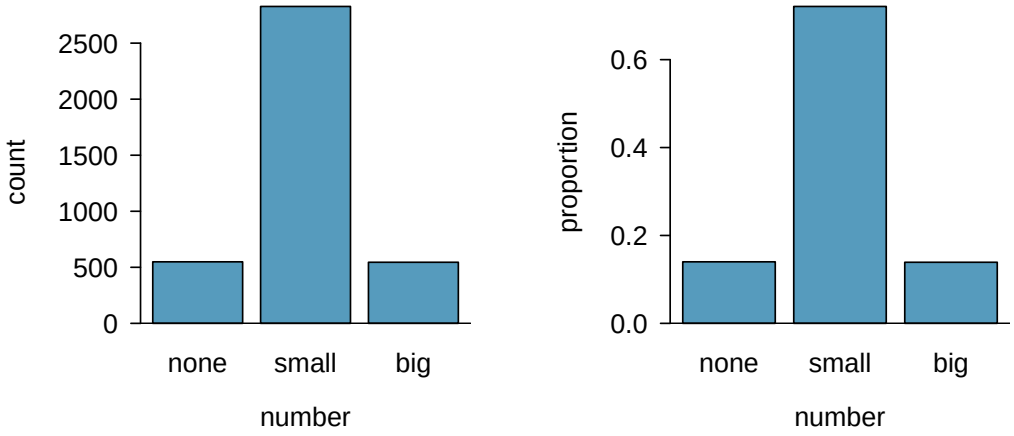


Figure 1.34: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

1.7.2 Row and column proportions

Table 1.35 shows the row proportions for Table 1.32. The **row proportions** are computed as the counts divided by their row totals. The value 149 at the intersection of `spam` and `none` is replaced by $149/367 = 0.406$, i.e. 149 divided by its row total, 367. So what does 0.406 represent? It corresponds to the proportion of spam emails in the sample that do not have any numbers.

	none	small	big	Total
spam	$149/367 = 0.406$	$168/367 = 0.458$	$50/367 = 0.136$	1.000
not spam	$400/3554 = 0.113$	$2657/3554 = 0.748$	$495/3554 = 0.139$	1.000
Total	$549/3921 = 0.140$	$2827/3921 = 0.721$	$545/3921 = 0.139$	1.000

Table 1.35: A contingency table with row proportions for the `spam` and `number` variables.

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. Table 1.36 shows such a table, and here the value 0.271 indicates that 27.1% of emails with no numbers were spam. This rate of spam is much higher than emails with only small numbers (5.9%) or big numbers (9.2%). Because these spam rates vary between the three levels of `number` (`none`, `small`, `big`), this provides evidence that the `spam` and `number` variables are associated.

	none	small	big	Total
spam	$149/549 = 0.271$	$168/2827 = 0.059$	$50/545 = 0.092$	$367/3921 = 0.094$
not spam	$400/549 = 0.729$	$2659/2827 = 0.941$	$495/545 = 0.908$	$3684/3921 = 0.906$
Total	1.000	1.000	1.000	1.000

Table 1.36: A contingency table with column proportions for the `spam` and `number` variables.

We could also have checked for an association between `spam` and `number` in Table 1.35 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of emails with no numbers, small numbers, and big numbers varied from `spam` to `not spam`.

Guided Practice 1.39 What does 0.458 represent in Table 1.35? What does 0.059 represent in Table 1.36?[42]

Guided Practice 1.40 What does 0.139 at the intersection of `not spam` and `big` represent in Table 1.35? What does 0.908 represent in the Table 1.36?[43]

Example 1.41 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 of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content. A contingency table for the `spam` and `format` variables from the `email` data set are shown in Table 1.37. Recall that an HTML email is an email with the capacity for special formatting, e.g. bold text. In Table 1.37, which would be more helpful to someone hoping to classify email as spam or regular email: row or column proportions?

Such a person 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, such as `number` and other variables, we stand a reasonable chance of being able to classify some email as spam or not spam. This is a topic we will return to in Chapter 6.

	text	HTML	Total
spam	209	158	367
not spam	986	2568	3554
Total	1195	2726	3921

Table 1.37: A contingency table for `spam` and `format`.

Example 1.41 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.

Guided Practice 1.42 Look back to Tables 1.35 and 1.36. Which would be more useful to someone hoping to identify spam emails using the `number` variable?[44]

[42] 0.458 represents the proportion of spam emails that had a small number. 0.058 represents the fraction of emails with small numbers that are spam.

[43] 0.139 represents the fraction of non-spam email that had a big number. 0.908 represents the fraction of emails with big numbers that are non-spam emails.

[44] The column proportions in Table 1.36 will probably be most useful, which makes it easier to see that emails with small numbers are spam about 5.9% of the time (relatively rare). We would also see that about 27.1% of emails with no numbers are spam, and 9.2% of emails with big numbers are spam.

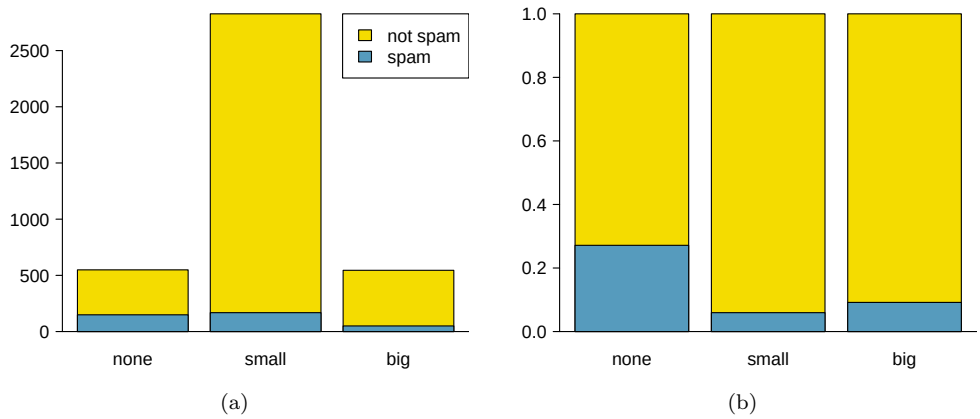


Figure 1.38: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by `spam`. (b) Standardized version of Figure (a).

1.7.3 Segmented bar and mosaic plots

Contingency tables using row or column proportions are especially useful for examining how two categorical variables are related. Segmented bar and mosaic plots provide a way to visualize the information in these tables.

A **segmented bar plot** is a graphical display of contingency table information. For example, a segmented bar plot representing Table 1.36 is shown in Figure 1.38(a), where we have first created a bar plot using the `number` variable and then separated each group by the levels of `spam`. The column proportions of Table 1.36 have been translated into a standardized segmented bar plot in Figure 1.38(b), which is a helpful visualization of the fraction of spam emails in each level of `number`.

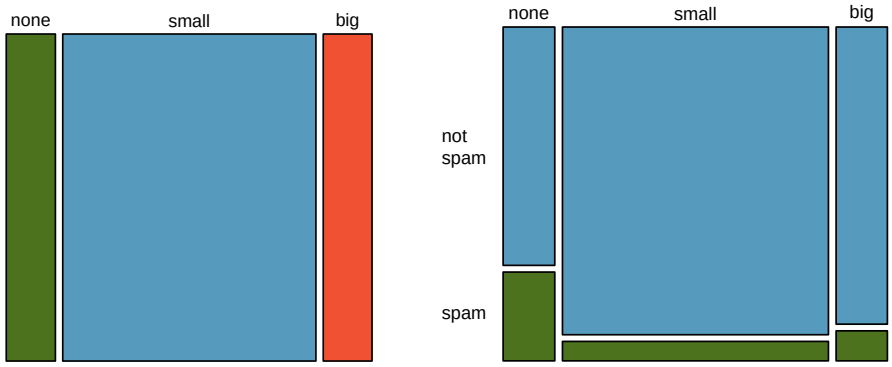
Example 1.43 Examine both of the segmented bar plots. Which is more useful?

Figure 1.38(a) contains more information, but Figure 1.38(b) presents the information more clearly. This second plot makes it clear that emails with no number have a relatively high rate of spam email – about 27%! On the other hand, less than 10% of email with small or big numbers are spam.

Since the proportion of spam changes across the groups in Figure 1.38(b), we can conclude the variables are dependent, which is something we were also able to discern using table proportions. Because both the `none` and `big` groups have relatively few observations compared to the `small` group, the association is more difficult to see in Figure 1.38(a).

In some other cases, a segmented bar plot that is not standardized will be more useful in communicating important information. Before settling on a particular segmented bar plot, create standardized and non-standardized forms and decide which is more effective at communicating features of the data.

A **mosaic plot** is a graphical display of contingency table information that is similar to a bar plot for one variable or a segmented bar plot when using two variables. Figure 1.39(a) shows a mosaic plot for the `number` variable. Each column represents a level of `number`, and the column widths correspond to the proportion of emails of each number type. For instance, there are fewer emails with no numbers than emails with only small numbers, so



(a) (b)

Figure 1.39: The one-variable mosaic plot for `number` and the two-variable mosaic plot for both `number` and `spam`.

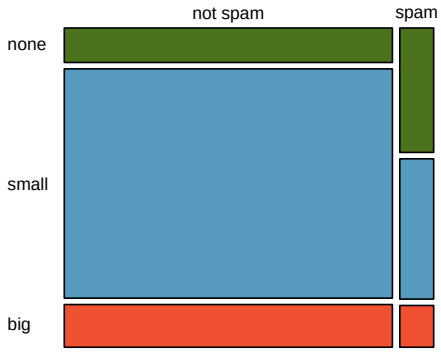


Figure 1.40: Mosaic plot where emails are grouped by the `number` variable after they've been divided into `spam` and `not spam`.

the no number email column is slimmer. In general, mosaic plots use box *areas* to represent the number of observations.

This one-variable mosaic plot is further divided into pieces in Figure 1.39(b) using the `spam` variable. Each column is split proportionally according to the fraction of emails that were spam in each number category. For example, the second column, representing emails with only small numbers, was divided into emails that were spam (lower) and not spam (upper). As another example, the bottom of the third column represents spam emails that had big numbers, and the upper part of the third column represents regular emails that had big numbers. We can again use this plot to see that the `spam` and `number` 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 version of the segmented bar plot.

In a similar way, a mosaic plot representing row proportions of Table 1.32 could be constructed, as shown in Figure 1.40. However, because it is more insightful for this application to consider the fraction of spam in each category of the `number` variable, we prefer Figure 1.39(b).

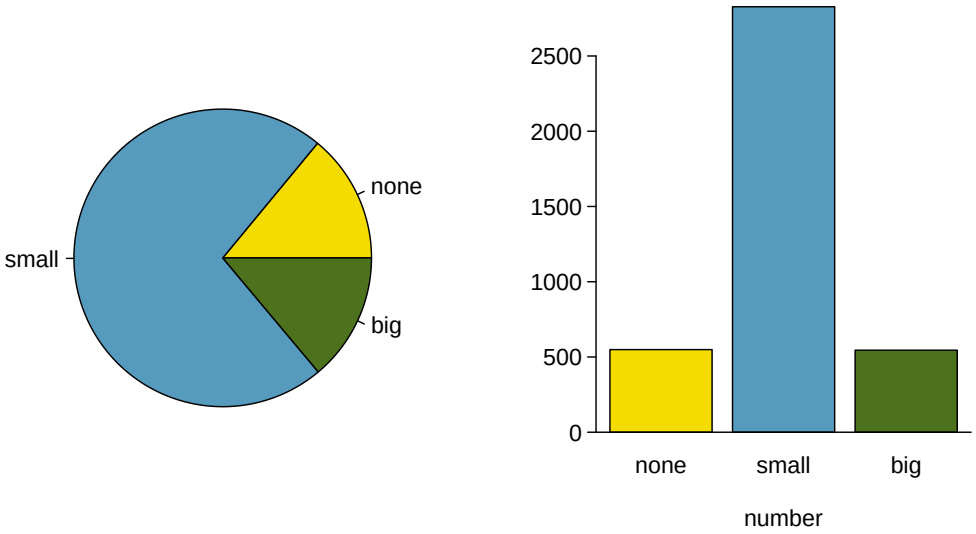


Figure 1.41: A pie chart and bar plot of `number` for the `email` data set.

1.7.4 The only pie chart you will see in this book

While pie charts are well known, they are not typically as useful as other charts in a data analysis. A **pie chart** is shown in Figure 1.41 alongside a bar plot. It is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions. In the case of the `none` and `big` categories, the difference is so slight you may be unable to distinguish any difference in group sizes for either plot!

1.7.5 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. Here two convenient methods are introduced: side-by-side box plots and hollow histograms.

We will take a look again at the `county` data set and compare the median household income for counties that gained population from 2000 to 2010 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 unjustified.

There were 2,041 counties where the population increased from 2000 to 2010, and there were 1,099 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 Table 1.42 to give a better sense of some of the raw 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 1.43, 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 1.43.

population gain						no gain		
41.2	33.1	30.4	37.3	79.1	34.5	40.3	33.5	34.8
22.9	39.9	31.4	45.1	50.6	59.4	29.5	31.8	41.3
47.9	36.4	42.2	43.2	31.8	36.9	28	39.1	42.8
50.1	27.3	37.5	53.5	26.1	57.2	38.1	39.5	22.3
57.4	42.6	40.6	48.8	28.1	29.4	43.3	37.5	47.1
43.8	26	33.8	35.7	38.5	42.3	43.7	36.7	36
41.3	40.5	68.3	31	46.7	30.5	35.8	38.7	39.8
68.3	48.3	38.7	62	37.6	32.2	46	42.3	48.2
42.6	53.6	50.7	35.1	30.6	56.8	38.6	31.9	31.1
66.4	41.4	34.3	38.9	37.3	41.7	37.6	29.3	30.1
51.9	83.3	46.3	48.4	40.8	42.6	57.5	32.6	31.1
44.5	34	48.7	45.2	34.7	32.2	46.2	26.5	40.1
39.4	38.6	40	57.3	45.2	33.1	38.4	46.7	25.9
43.8	71.7	45.1	32.2	63.3	54.7	36.4	41.5	45.7
71.3	36.3	36.4	41	37	66.7	39.7	37	37.7
50.2	45.8	45.7	60.2	53.1		21.4	29.3	50.1
35.8	40.4	51.5	66.4	36.1		43.6	39.8	

Table 1.42: In this table, median household income (in $1000s) from a random sample of 100 counties that gained population over 2000-2010 are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

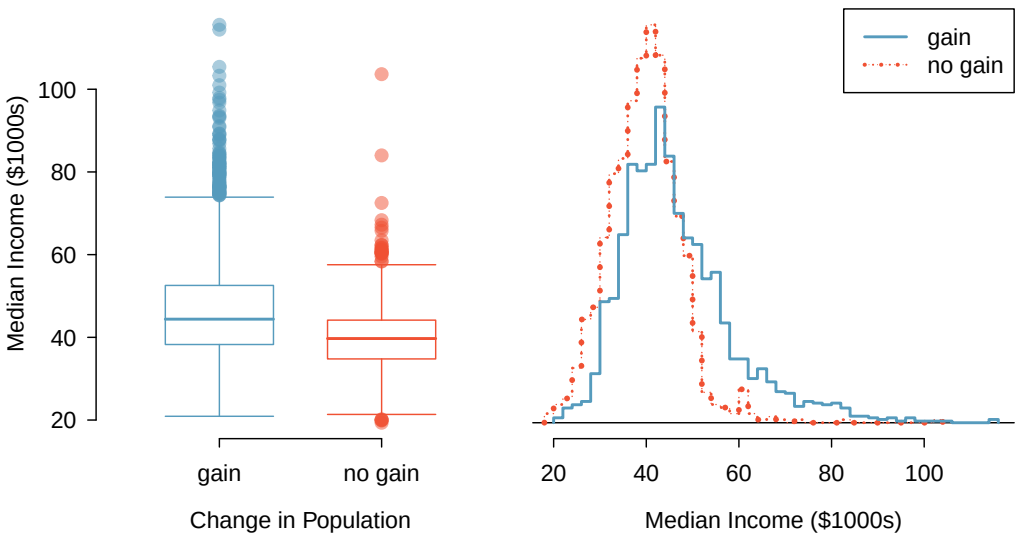


Figure 1.43: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.

⊙ **Guided Practice 1.44** Use the plots in Figure 1.43 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?[45]

⊙ **Guided Practice 1.45** What components of each plot in Figure 1.43 do you find most useful?[46]

[45] 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. There is a secondary small bump at about $60,000 for the *no gain* group, visible in the hollow histogram plot, that seems out of place. (Looking into the data set, we would find that 8 of these 15 counties are in Alaska and Texas.) 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 using such a large data set.

[46] Answers will vary. The side-by-side 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.

1.8 Exercises

1.8.1 Case study

1.1 Migraine and acupuncture. 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 nonacupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[47]

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

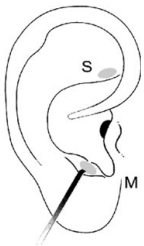


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

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

(b) At first glance, does acupuncture appear to be an effective treatment for migraines? Explain your reasoning.

(c) Do the data provide convincing evidence that there is a real pain reduction for those patients in the treatment group? Or do you think that the observed difference might just be due to chance?

1.2 Sinusitis and antibiotics, Part I. 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 significant improvement in symptoms. The distribution of responses are summarized below.[48]

		Self-reported significant improvement in symptoms		
		Yes	No	Total
Group	Treatment	66	19	85
	Control	65	16	81
	Total	131	35	166

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

(b) Based on your findings in part (a), which treatment appears to be more effective for sinusitis?

(c) Do the data provide convincing evidence that there is a difference in the improvement rates of sinusitis symptoms? Or do you think that the observed difference might just be due to chance?

[47] 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.

[48] 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.

1.8.2 Data basics

1.3 Identify study components, Part I. Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question in the studies described below.

(a) Researchers collected data to examine the relationship between 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.[49]

(b) 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 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.[50]

1.4 Identify study components, Part II. Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question of the studies described below.

(a) While obesity is measured based on body fat percentage (more than 35% body fat for women and more than 25% for men), precisely measuring body fat percentage is difficult. Body mass index (BMI), calculated as the ratio $weight/height^2$, is often used as an alternative indicator for obesity. A common criticism of BMI is that it assumes the same relative body fat percentage regardless of age, sex, or ethnicity. In order to determine how useful BMI is for predicting body fat percentage across age, sex and ethnic groups, researchers studied 202 black and 504 white adults who resided in or near New York City, were ages 20-94 years old, had BMIs of 18-35 kg/m^2, and who volunteered to be a part of the study. Participants reported their age, sex, and ethnicity and were measured for weight and height. Body fat percentage was measured by submerging the participants in water.[51]

(b) 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 they were for children in a nearby laboratory, but that they could take some if they wanted. Participants completed unrelated tasks and then reported the number of candies they had taken. It was found that those in the upper-class rank condition took more candy than did those in the lower-rank condition.[52]

[49] 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.

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

[51] Gallagher et al. "How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups?" In: *American Journal of Epidemiology* 143.3 (1996), pp. 228–239.

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

1.5 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.[53]

(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).

1.6 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.[54]

	gender	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.

1.8.3 Overview of data collection principles

1.7 Generalizability and causality, Part I. Identify the population of interest and the sample in the studies described in Exercise 1.3. Also 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.

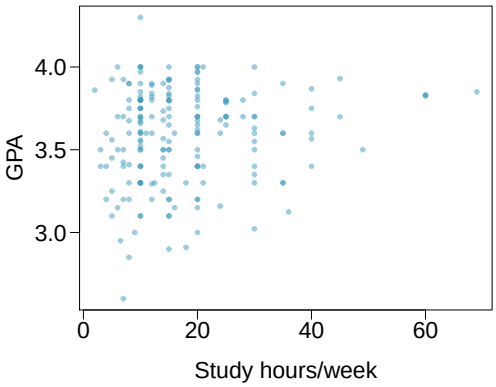
1.8 Generalizability and causality, Part II. Identify the population of interest and the sample in the studies described in Exercise 1.4. Also comment on whether or not the results of the study can be generalized to the population and if the findings of the study can be used to establish causal relationships.

[53] Photo by rtclauss on Flickr, Iris.; R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[54] Stats4Schools, Smoking.

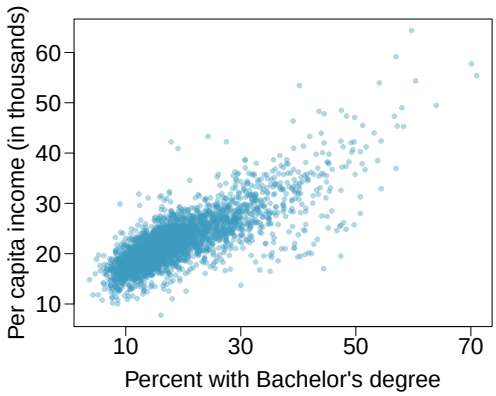
1.9 GPA and study time. A survey was conducted on 218 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about their GPA and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.

(a) What is the explanatory variable and what is the response variable?
(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
(c) Is this an experiment or an observational study?
(d) Can we conclude that studying longer hours leads to higher GPAs?



1.10 Income and education. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

(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?



1.8.4 Observational studies and sampling strategies

1.11 Propose a sampling strategy. 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?

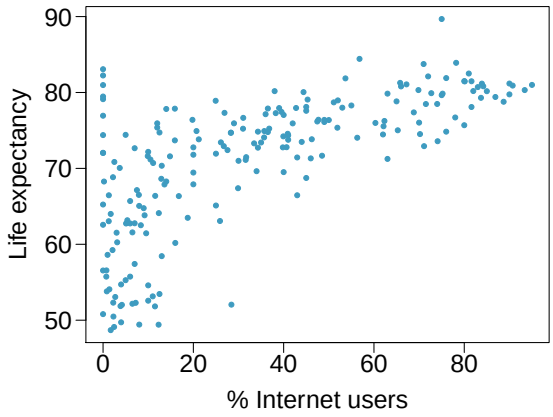
(b) Suggest a sampling strategy for carrying out this study.

1.12 Internet use and life expectancy. The scatterplot below shows the relationship between estimated life expectancy at birth as of 2012[55] and percentage of internet users in 2010[56] in 208 countries.

(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.



1.13 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.

1.14 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. Three 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, and makes sure each one of them fills out the survey.

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

1.15 Family size. Suppose we want to estimate family size, where family is defined as one or more parents living with children. If we select students at random at an elementary school and ask them what their family size is, will our average be biased? If so, will it overestimate or underestimate the true value?

[55] CIA Factbook, Country Comparison: Life Expectancy at Birth, 2012.

[56] ITU World Telecommunication/ICT Indicators database, World Telecommunication/ICT Indicators Database, 2012.

1.16 Flawed reasoning. Identify the flaw 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 required 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) A 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.

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

(a) An article called *Risks: Smokers Found More Prone to Dementia* states the following:[57]

> "Researchers analyzed the data of 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50 to 60 years old. Twenty-three years later, about one-quarter of the group, or 5,367, had dementia, including 1,136 with Alzheimers disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37 percent more likely than nonsmokers to develop dementia, and the risks went up sharply with increased smoking; 44 percent 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 called *The School Bully Is Sleepy* states the following:[58]

> "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?

1.18 Shyness on Facebook. Given the anonymity afforded to individuals in online interactions, researchers hypothesized that shy individuals would have more favorable attitudes toward Facebook and that shyness would be positively correlated with time spent on Facebook. They also hypothesized that shy individuals would have fewer Facebook "Friends" just like they have fewer friends than non-shy individuals have in the offline world. Data were collected on 103 undergraduate students at a university in southwestern Ontario via online questionnaires. The study states "Participants were recruited through the university's psychology participation pool. After indicating an interest in the study, participants were sent an e-mail containing the study's URL as well as the necessary login credentials." Are the results of this study generalizable to the population of all Facebook users?[59]

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

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

[59] E.S. Orr et al. "The influence of shyness on the use of Facebook in an undergraduate sample". In: *CyberPsychology & Behavior* 12.3 (2009), pp. 337–340.

1.8.5 Experiments

1.19 Vitamin supplements. In order 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.[60]

(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.

1.20 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.

1.21 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?

[60] 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.

1.22 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 evenly 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.[61]

(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.

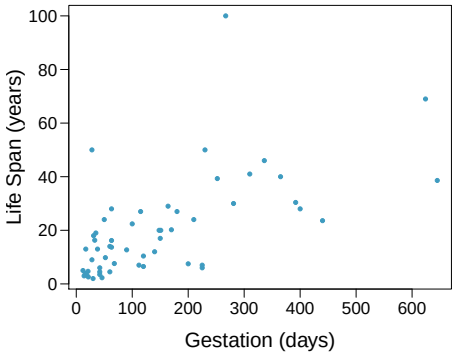
1.8.6 Examining numerical data

1.23 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.[62]

(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.



1.24 Office productivity. Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.

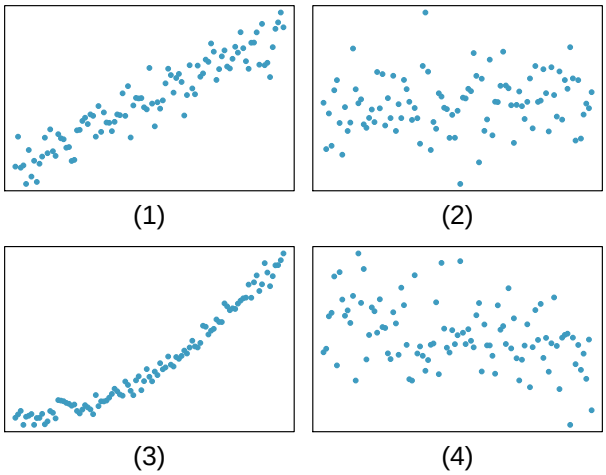
[61] D.C. Nieman et al. "Chia seed does not promote weight loss or alter disease risk factors in overweight adults". In: *Nutrition Research* 29.6 (2009), pp. 414–418.

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

1.25 Associations. Indicate which of the plots show a

(a) positive association
(b) negative association
(c) no association

Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.



1.26 Parameters and statistics. Identify which value represents the sample mean and which value represents the claimed population mean.

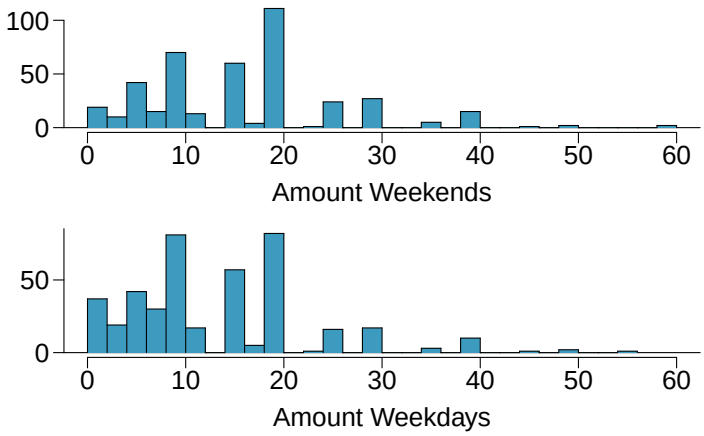
(a) A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night.
(b) American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.
(c) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2012.

1.27 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?

1.28 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?

1.29 Smoking habits of UK residents, Part I. Exercise 1.6 introduces a data set on the smoking habits of UK residents. Below are histograms displaying the distributions of the number of cigarettes smoked on weekdays and weekends, excluding non-smokers. Describe the two distributions and compare them.



1.30 Stats scores. Below are the final scores of 20 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.

1.31 Smoking habits of UK residents, Part II. A random sample of 5 smokers from the data set discussed in Exercises 1.6 and 1.29 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?

1.32 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.

1.33 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, 8, 7, 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

1.34 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

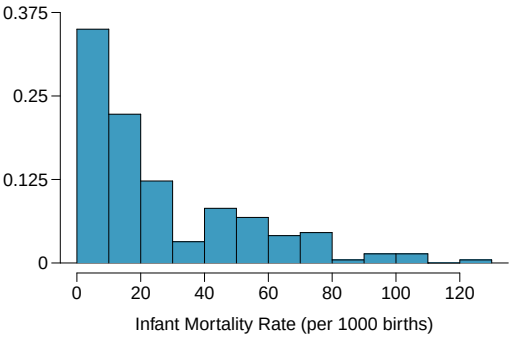
1.35 Box plot. Create a box plot for the data given in Exercise 1.30. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

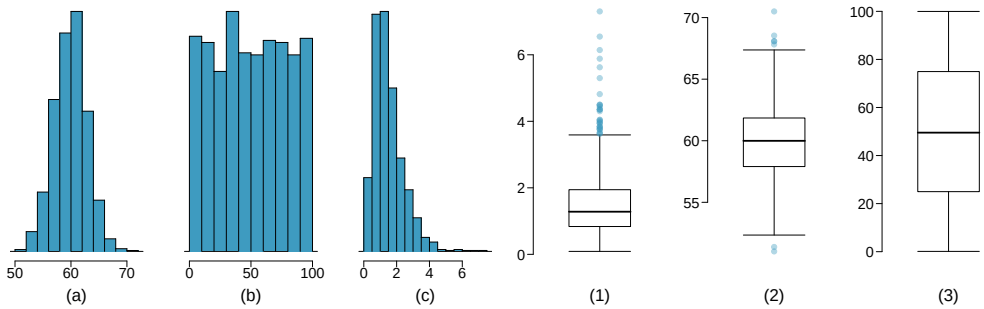
1.36 Infant mortality. The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates in 2012 for 222 countries.[63]

(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.

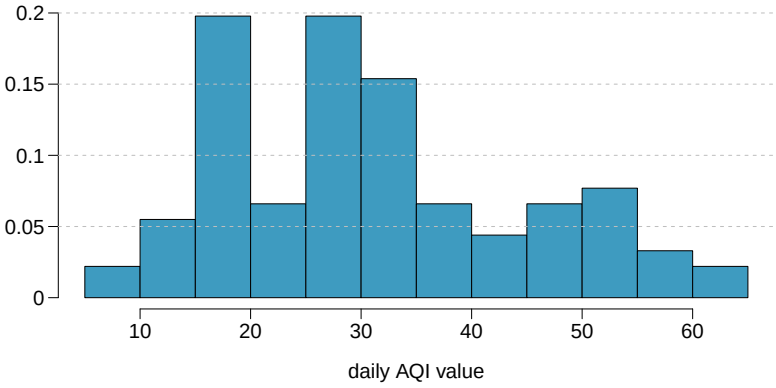


1.37 Matching histograms and box plots. Describe the distribution in the histograms below and match them to the box plots.



[63] CIA Factbook, Country Comparison: Infant Mortality Rate, 2012.

1.38 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environmental 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.[64]

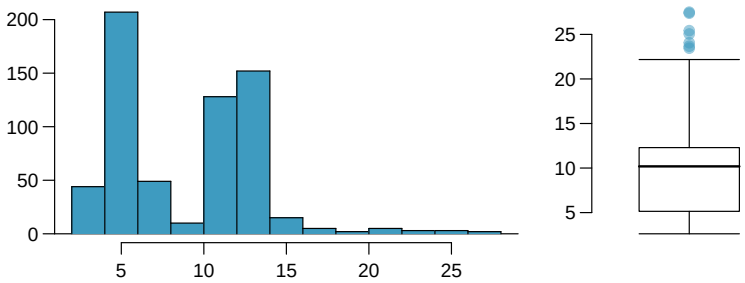


(a) Estimate the median AQI value of this sample.

(b) Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.

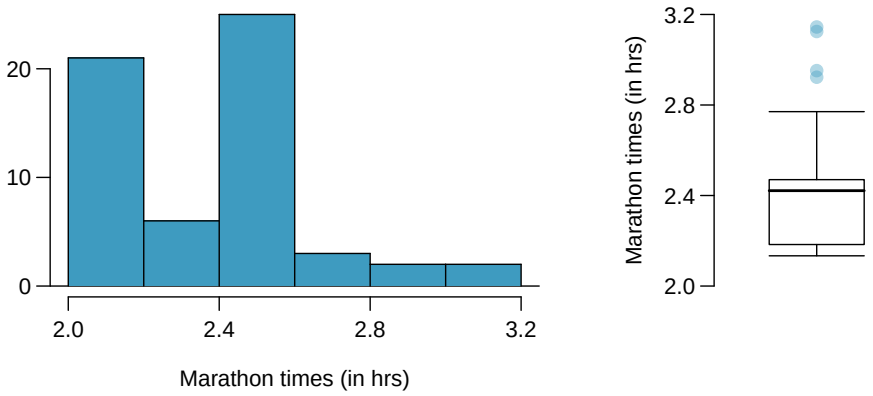
(c) Estimate Q1, Q3, and IQR for the distribution.

1.39 Histograms and 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?



[64] US Environmental Protection Agency, AirData, 2011.

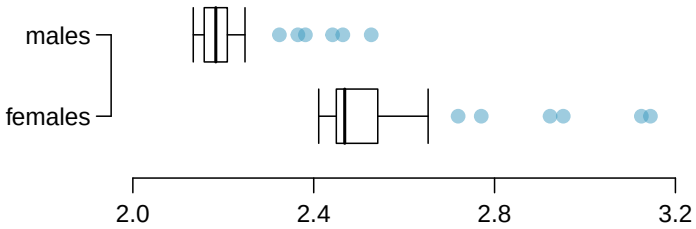
1.40 Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



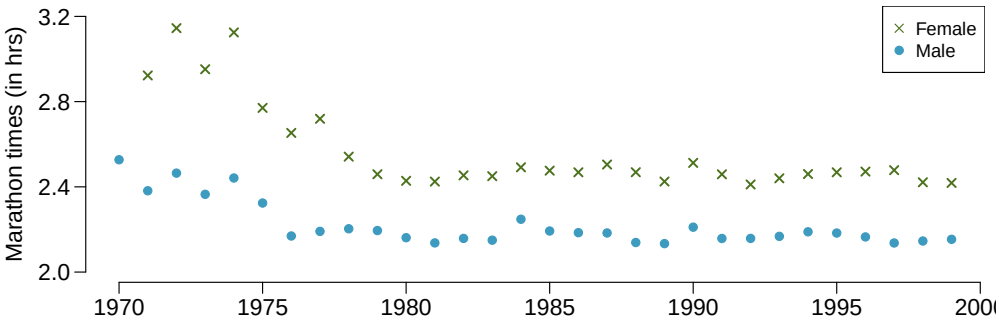
(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

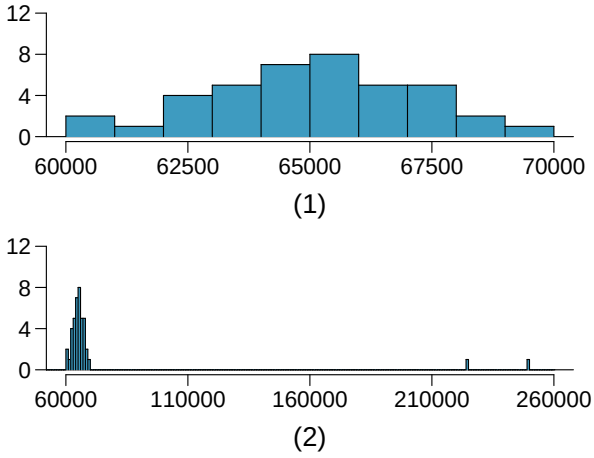
(c) Compare the distribution of marathon times for men and women based on the box plot shown below.



(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.



1.41 Robust statistics. 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)
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?

1.42 Distributions and appropriate statistics. For each of the following, describe 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.

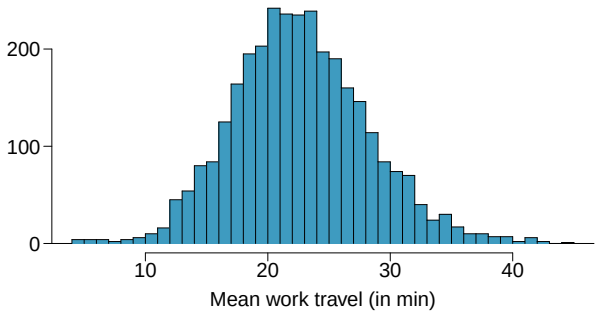
(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(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.

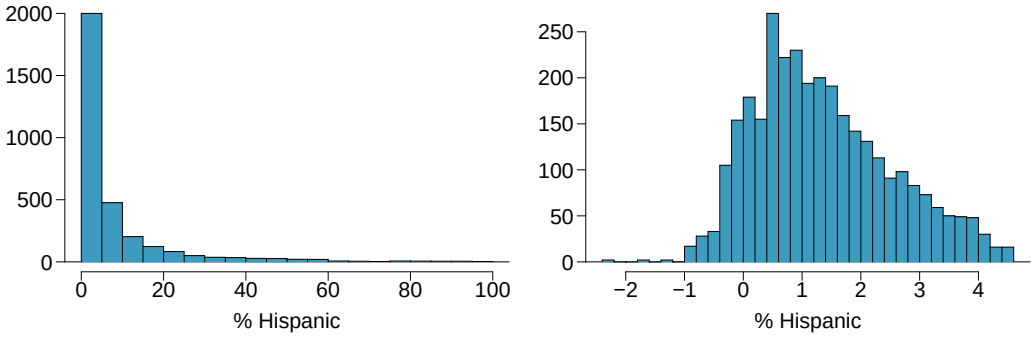
(c) Number of alcoholic drinks consumed by college students in a given week.

(d) Annual salaries of the employees at a Fortune 500 company.

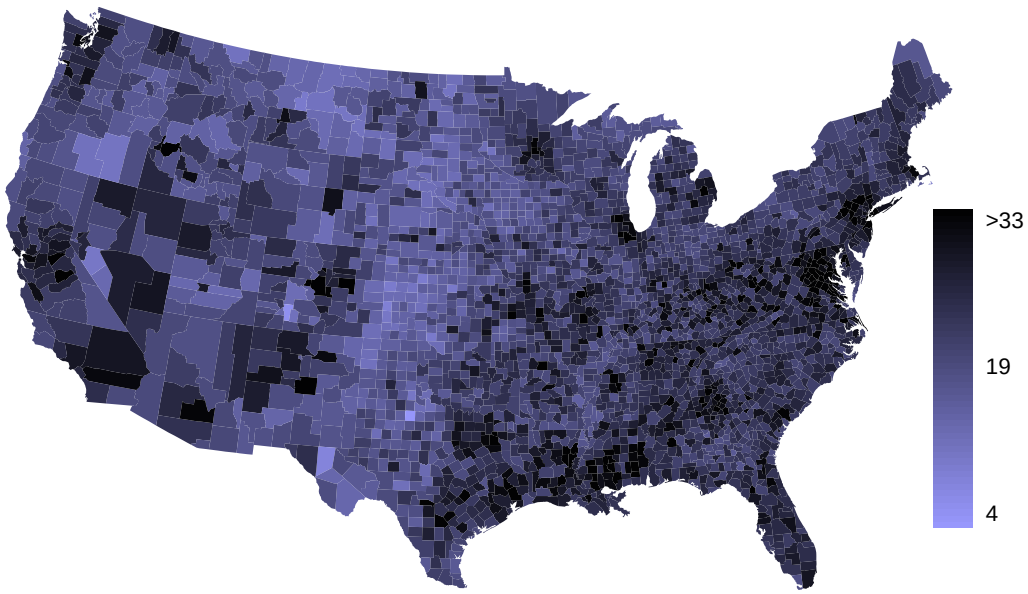
1.43 Commuting times, Part I. The histogram to the right shows the distribution of mean commuting times in 3,143 US counties in 2010. Describe the distribution and comment on whether or not a log transformation may be advisable for these data.



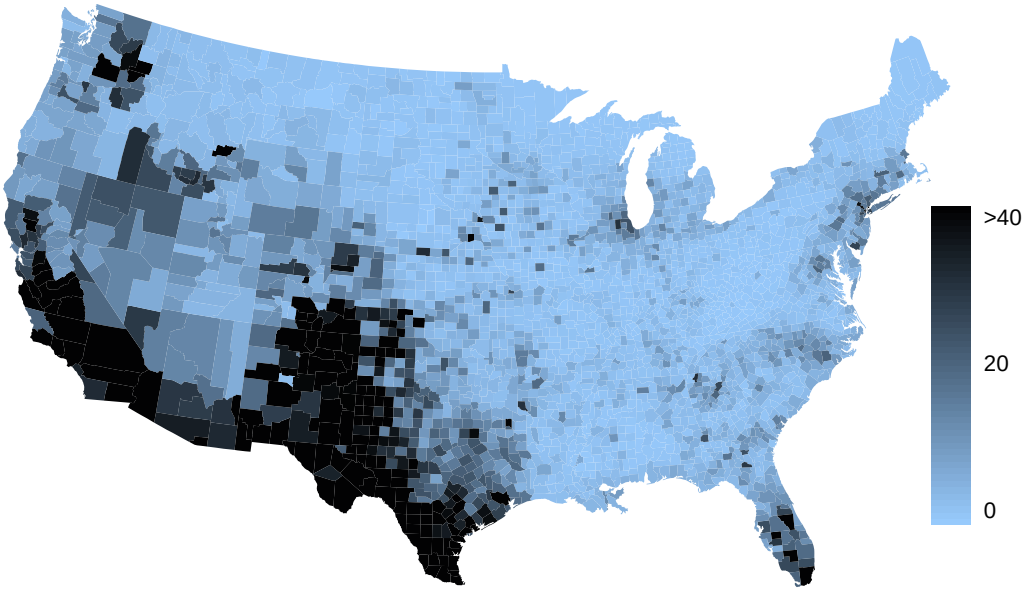
1.44 Hispanic population, Part I. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010. Also shown is a histogram of logs of these values. Describe the distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.



1.45 Commuting times, Part II. Exercise 1.43 displays histograms of mean commuting times in 3,143 US counties in 2010. Describe the spatial distribution of commuting times using the map below.



1.46 Hispanic population, Part II. Exercise 1.44 displays histograms of the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010.



(a) What features of this distribution are apparent in the map but not in the histogram?

(b) 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.

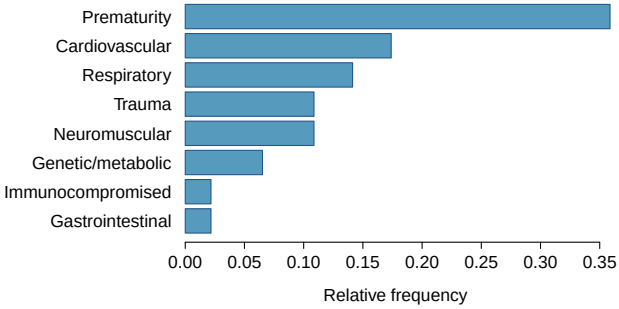
1.8.7 Considering categorical data

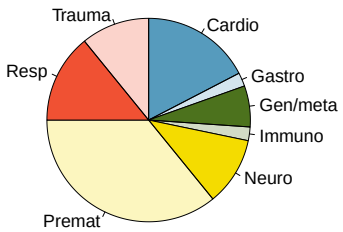
1.47 Antibiotic use in children. The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.

(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?





1.48 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.[65]

		Political ideology			
		Conservative	Moderate	Liberal	Total
Response	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
	(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?

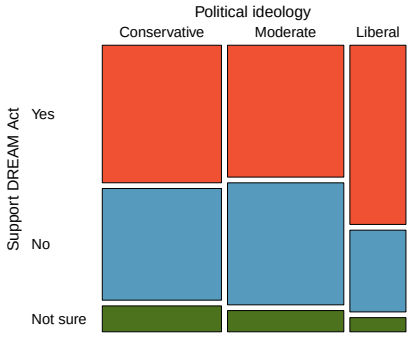
(b) What percent of these Tampa, FL voters are in favor of the citizenship option?

(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 and liberal share this view?

(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

1.49 Views on the DREAM Act. The same survey from Exercise 1.48 also asked respondents 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. Based on the mosaic plot shown on the right, are views on the DREAM Act and political ideology independent?

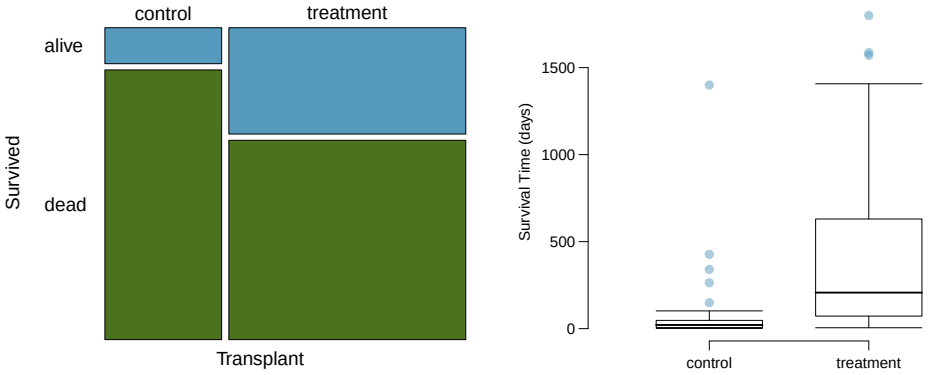


[65] SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

1.50 Heart transplants, Part I. 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. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study.[66]

(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 suggest about the efficacy (effectiveness) of transplants?



[66] B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

Chapter 2

Foundation for inference

● **Example 2.1** 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, they are probably not exactly the same. We would probably observe a small difference due to chance.

⊙ **Guided Practice 2.2** 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?[1]

Studying randomness of this form is a key focus of statistics. In this chapter, we'll explore this type of randomness in the context of several applications, and we'll learn new tools and ideas that will be applied throughout the rest of the book.

2.1 Randomization case study: gender discrimination

We consider a study investigating gender discrimination in the 1970s, which is set in the context of personnel decisions within a bank.[2] The research question we hope to answer is, "Are females discriminated against in promotion decisions made by male managers?"

2.1.1 Variability within data

The participants in this study were 48 male bank supervisors attending a management institute at the University of North Carolina in 1972. They were asked to assume the role of the personnel director of a bank and were given a personnel file to judge whether the person should be promoted to a branch manager position. The files given to the participants were identical, except that half of them indicated the candidate was male and the other half indicated the candidate was female. These files were randomly assigned to the subjects.

[1] We would be assuming that these two variables are *independent*, meaning they are unrelated.

[2] Rosen B and Jerdee T. 1974. Influence of sex role stereotypes on personnel decisions. Journal of Applied Psychology 59(1):9-14.

⊙ **Guided Practice 2.3** Is this an observational study or an experiment? How does the type of study impact what can be inferred from the results?[3]

For each supervisor we recorded the gender associated with the assigned file and the promotion decision. Using the results of the study summarized in Table 2.1, we would like to evaluate if females are unfairly discriminated against in promotion decisions. In this study, a smaller proportion of females are promoted than males (0.583 versus 0.875), but it is unclear whether the difference provides *convincing evidence* that females are unfairly discriminated against.

		`decision`		
		promoted	not promoted	Total
`gender`	male	21	3	24
	female	14	10	24
	Total	35	13	48

Table 2.1: Summary results for the gender discrimination study.

● **Example 2.4** Statisticians are sometimes called upon to evaluate the strength of evidence. When looking at the rates of promotion for males and females in this study, why might we be tempted to immediately conclude that females are being discriminated against?

The large difference in promotion rates (58.3% for females versus 87.5% for males) suggest there might be discrimination against women in promotion decisions. However, we cannot yet be sure if the observed difference represents discrimination 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 promotion decisions were independent of gender.

Example 2.4 is a reminder that the observed outcomes in the sample may not perfectly reflect the true relationships between variables in the underlying population. Table 2.1 shows there were 7 fewer promotions in the female group than in the male group, a difference in promotion rates of 29.2% $\left(\frac{21}{24} - \frac{14}{24} = 0.292\right)$. This observed difference is what we call a **point estimate** of the true effect. The point estimate of the difference is large, but the sample size for the study is small, making it unclear if this observed difference represents discrimination or whether it is simply due to chance. We label these two competing claims, H_0 and H_A:

H_0: **Null hypothesis.** The variables `gender` and `decision` are independent. They have no relationship, and the observed difference between the proportion of males and females who were promoted, 29.2%, was due to chance.

H_A: **Alternative hypothesis.** The variables `gender` and `decision` are *not* independent. The difference in promotion rates of 29.2% was not due to chance, and equally qualified females are less likely to be promoted than males.

[3]The study is an experiment, as subjects were randomly assigned a male file or a female file. Since this is an experiment, the results can be used to evaluate a causal relationship between gender of a candidate and the promotion decision.

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.

What would it mean if the null hypothesis, which says the variables `gender` and `decision` are unrelated, is true? It would mean each banker would decide whether to promote the candidate without regard to the gender indicated on the file. That is, the difference in the promotion percentages would be due to the way the files were randomly divided to the bankers, and the randomization just happened to give rise to a relatively large difference of 29.2%.

Consider the alternative hypothesis: bankers were influenced by which gender was listed on the personnel file. If this was true, and especially if this influence was substantial, we would expect to see some difference in the promotion rates of male and female candidates. If this gender bias was against females, we would expect a smaller fraction of promotion recommendations for female personnel files relative to the male files.

We will choose between these two competing claims by assessing if the data conflict so much with H_0 that the null hypothesis 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 that these data provide strong evidence of discrimination.

2.1.2 Simulating the study

Table 2.1 shows that 35 bank supervisors recommended promotion and 13 did not. Now, suppose the bankers' decisions were independent of gender. Then, if we conducted the experiment again with a different random assignment of files, differences in promotion rates would be based only on random fluctuation. We can actually perform this **randomization**, which simulates what would have happened if the bankers' decisions had been independent of gender but we had distributed the files differently.[4]

In this **simulation**, we thoroughly shuffle 48 personnel files, 24 labeled `male` and 24 labeled `female`, and deal these files into two stacks. We will deal 35 files into the first stack, which will represent the 35 supervisors who recommended promotion. The second stack will have 13 files, and it will represent the 13 supervisors who recommended against promotion. Then, as we did with the original data, we tabulate the results and determine the fraction of `male` and `female` who were promoted.

Since the randomization of files in this simulation is independent of the promotion decisions, any difference in the two fractions is entirely due to chance. Table 2.2 show the results of such a simulation.

⊙ **Guided Practice 2.5** What is the difference in promotion rates between the two simulated groups in Table 2.2? How does this compare to the observed difference 29.2% from the actual study?[5]

[4] The test procedure we employ in this section is formally called a **permutation test**.

[5] $18/24 - 17/24 = 0.042$ or about 4.2% in favor of the men. This difference due to chance is much smaller than the difference observed in the actual groups.

		decision		
		promoted	not promoted	Total
gender_simulated	male	18	6	24
	female	17	7	24
	Total	35	13	48

Table 2.2: Simulation results, where any difference in promotion rates between `male` and `female` is purely due to chance.

2.1.3 Checking for independence

We computed one possible difference under the null hypothesis in Guided Practice 2.5, which represents one difference due to chance. While in this first simulation, we physically dealt out files, it is much more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance: -0.042. And another: 0.208. 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 2.3 shows a plot of the differences found from 100 simulations, where each dot represents a simulated difference between the proportions of male and female files recommended for promotion.

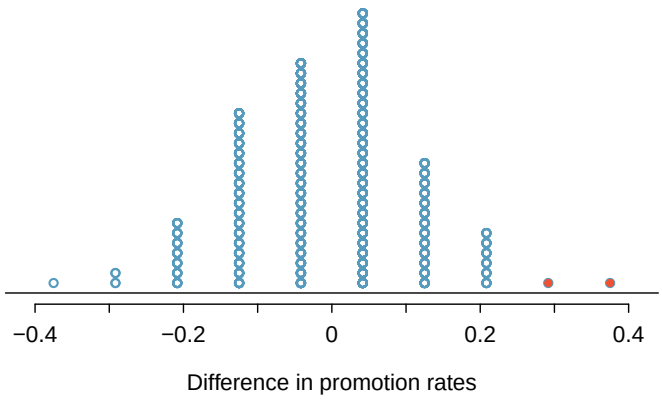


Figure 2.3: A stacked dot plot of differences from 100 simulations produced under the null hypothesis, H_0, where `gender_simulated` and `decision` are independent. Two of the 100 simulations had a difference of at least 29.2%, the difference observed in the study, and are shown as solid dots.

Note that the distribution of these simulated differences is centered around 0. Because we simulated differences in a way that made no distinction between men and women, this makes sense: we should expect differences from chance alone to fall around zero with some random fluctuation for each simulation.

Example 2.6 How often would you observe a difference of at least 29.2% (0.292) according to Figure 2.3? Often, sometimes, rarely, or never?

It appears that a difference of at least 29.2% due to chance alone would only happen about 2% of the time according to Figure 2.3. Such a low probability indicates that observing such a large difference from chance is rare.

The difference of 29.2% is a rare event if there really is no impact from listing gender in the candidates' files, which provides us with two possible interpretations of the study results:

H_0: **Null hypothesis.** Gender has no effect on promotion decision, and we observed a difference that is so large that it would only happen rarely.

H_A: **Alternative hypothesis.** Gender has an effect on promotion decision, and what we observed was actually due to equally qualified women being discriminated against in promotion decisions, which explains the large difference of 29.2%.

When we conduct formal studies, we reject a skeptical position if the data strongly conflict with that position.[6] In our analysis, we determined that there was only a ≈2% probability of obtaining a sample where ≥29.2% more males than females get promoted by chance alone, so we conclude the data provide strong evidence of gender discrimination against women by the supervisors. In this case, we reject the null hypothesis in favor of the alternative.

Statistical inference is the practice of making decisions and conclusions from data in the context of uncertainty. Errors do occur, just like rare events, and the data set at hand might lead us to the wrong 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. Before getting into the nuances of hypothesis testing, let's work through another case study.

2.2 Randomization case study: opportunity cost

How rational and consistent is the behavior of the typical American college student? In this section, we'll explore whether college student consumers always consider an obvious fact: money not spent now can be spent later.

In particular, we are interested in whether reminding students about this well-known fact about money causes them to be a little thriftier. A skeptic might think that such a reminder would have no impact. We can summarize these two perspectives using the null and alternative hypothesis framework.

H_0: **Null hypothesis.** Reminding students that they can save money for later purchases will not have any impact on students' spending decisions.

H_A: **Alternative hypothesis.** Reminding students that they can save money for later purchases will reduce the chance they will continue with a purchase.

In this section, we'll explore an experiment conducted by researchers that investigates this very question for students at a university in the southwestern United States.[7]

[6] This reasoning does not generally extend to anecdotal observations. Each of us observes incredibly rare events every day, events we could not possibly hope to predict. However, in the non-rigorous setting of anecdotal evidence, almost anything may appear to be a rare event, so the idea of looking for rare events in day-to-day activities is treacherous. For example, we might look at the lottery: there was only a 1 in 176 million chance that the Mega Millions numbers for the largest jackpot in history (March 30, 2012) would be (2, 4, 23, 38, 46) with a Mega ball of (23), but nonetheless those numbers came up! However, no matter what numbers had turned up, they would have had the same incredibly rare odds. That is, *any set of numbers we could have observed would ultimately be incredibly rare*. This type of situation is typical of our daily lives: each possible event in itself seems incredibly rare, but if we consider every alternative, those outcomes are also incredibly rare. We should be cautious not to misinterpret such anecdotal evidence.

[7] Frederick S, Novemsky N, Wang J, Dhar R, Nowlis S. 2009. Opportunity Cost Neglect. Journal of Consumer Research 36: 553-561.

2.2.1 Exploring the data set before the analysis

One-hundred and fifty students were recruited for the study, and each was given the following statement:

> Imagine that you have been saving some extra money on the side to make some purchases, and on your most recent visit to the video store you come across a special sale on a new video. This video is one with your favorite actor or actress, and your favorite type of movie (such as a comedy, drama, thriller, etc.). This particular video that you are considering is one you have been thinking about buying for a long time. It is available for a special sale price of $14.99.
>
> What would you do in this situation? Please circle one of the options below.

Half of the 150 students were randomized into a control group and were given the following two options:

(A) Buy this entertaining video.

(B) Not buy this entertaining video.

The remaining 75 students were placed in the treatment group, and they saw a slightly modified option (B):

(A) Buy this entertaining video.

(B) Not buy this entertaining video. Keep the $14.99 for other purchases.

Would the extra statement reminding students of an obvious fact impact the purchasing decision? Table 2.4 summarizes the study results.

	decision		
	buy DVD	not buy DVD	Total
control group	56	19	75
treatment group	41	34	75
Total	97	53	150

Table 2.4: Summary of student choices in the opportunity cost study.

It might be a little easier to review the results using row proportions, specifically considering the proportion of participants in each group who said they would buy or not buy the DVD. These summaries are given in Table 2.5.

	decision		
	buy DVD	not buy DVD	Total
control group	0.747	0.253	1.000
treatment group	0.547	0.453	1.000
Total	0.647	0.353	1.000

Table 2.5: The data from Table 2.4 summarized using row proportions. Row proportions are particularly useful here since we can view the proportion of *buy* and *not buy* decisions in each group.

We will define a **success** in this study as a student who chooses not to buy the DVD.[8] Then, the value of interest is the change in DVD purchase rates that results by reminding students that not spending money now means they can spend the money later. We can construct a point estimate for this difference as

$$\hat{p}_{trmt} - \hat{p}_{ctrl} = \frac{34}{75} - \frac{19}{75} = 0.453 - 0.253 = 0.200$$

The proportion of students who chose not to buy the DVD was 20% higher in the treatment group than the control group. However, is this result **statistically significant**? In other words, is a 20% difference between the two groups so prominent that it is unlikely to have occurred from chance alone?

2.2.2 Results from chance alone

The primary goal in this data analysis is to understand what sort of differences we might see if the null hypothesis were true, i.e. the treatment had no effect on students. For this, we'll use the same procedure we applied in Section 2.1: randomization.

Let's think about the data in the context of the hypotheses. If the null hypothesis (H_0) was true and the treatment had no impact on student decisions, then the observed difference between the two groups of 20% could be attributed entirely to chance. If, on the other hand, the alternative hypothesis (H_A) is true, then the difference indicates that reminding students about saving for later purchases actually impacts their buying decisions.

Just like with the gender discrimination study, we can perform a statistical analysis. Using the same randomization technique from the last section, let's see what happens when we simulate the experiment under the scenario where there is no effect from the treatment.

While we would in reality do this simulation on a computer, it might be useful to think about how we would go about carrying out the simulation without a computer. We start with 150 index cards and label each card to indicate the distribution of our response variable: decision. That is, 53 cards will be labeled "not buy DVD" to represent the 53 students who opted not to buy, and 97 will be labeled "buy DVD" for the other 97 students. Then we shuffle these cards throughly and divide them into two stacks of size 75, representing the simulated treatment and control groups. Any observed difference between the proportions of "not buy DVD" cards (what we earlier defined as *success*) can be attributed entirely to chance.

Example 2.7 If we are randomly assigning the cards into the simulated treatment and control groups, how many "not buy DVD" cards would we expect to end up with in each simulated group? What would be the expected difference between the proportions of "not buy DVD" cards in each group?

Answer: Since the simulated groups are of equal size, we would expect $53/2 = 26.5$, i.e. 26 or 27, "not buy DVD" cards in each simulated group, yielding a simulated point estimate of 0%. However, due to random fluctuations, we might actually observe a number a little above or below 26 and 27.

[8] Success is often defined in a study as the outcome of interest, and a "success" may or may not actually be a positive outcome. For example, researchers working on a study on HIV prevalence might define a "success" in the statistical sense as a patient who is HIV+. A more complete discussion of the term *success* will be given in Chapter 3.

The results of a randomization from chance alone is shown in Table 2.6. From this table, we can compute a difference that occurred from chance alone:

$$\hat{p}_{trmt,simulated} - \hat{p}_{ctrl,simulated} = \frac{24}{75} - \frac{29}{75} = 0.32 - 0.387 = -0.067$$

	decision		
	buy DVD	not buy DVD	Total
simulated-control group	46	29	75
simulated-treatment group	51	24	75
Total	97	53	150

Table 2.6: Summary of student choices against their simulated groups. The group assignment had no connection to the student decisions, so any difference between the two groups is due to chance.

Just one simulation will not be enough to get a sense of what sorts of differences would happen from chance alone. We'll simulate another set of simulated groups and compute the new difference: 0.013. And again: 0.067. And again: -0.173. We'll do this 1,000 times. The results are summarized in a dot plot in Figure 2.7, where each point represents a simulation. Since there are so many points, it is more convenient to summarize the results in a histogram such as the one in Figure 2.8, where the height of each histogram bar represents the fraction of observations in that group.

If there was no treatment effect, then we'd only observe a difference of at least +20% about 0.6% of the time, or about 1-in-150 times. That is really rare! Instead, we will conclude the data provide strong evidence there is a treatment effect: reminding students before a purchase that they could instead spend the money later on something else lowers the chance that they will continue with the purchase. Notice that we are able to make a causal statement for this study since the study is an experiment.

2.3 Hypothesis testing

In the last two sections, we utilized a **hypothesis test**, which is a formal technique for evaluating two competing possibilities. In each scenario, we described a **null hypothesis**, which represented either a skeptical perspective or a perspective of no difference. We also laid out an **alternative hypothesis**, which represented 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** (H_0) often represents either a skeptical perspective or a claim to be tested. The **alternative hypothesis** (H_A) represents an alternative claim under consideration and is often represented by a range of possible values for the value of interest.

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.

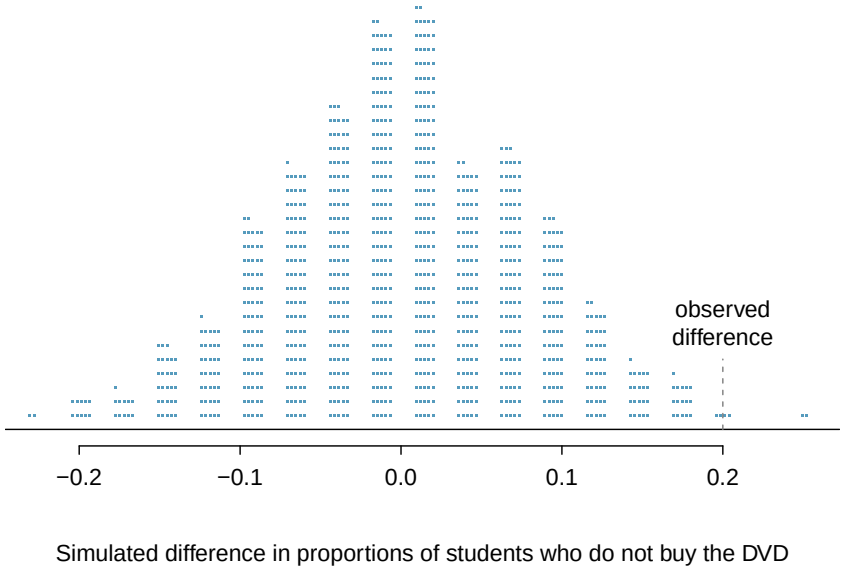


Figure 2.7: A stacked dot plot of 1,000 chance differences produced under the null hypothesis, H_0. Six of the 1,000 simulations had a difference of at least 20%, which was the difference observed in the study.

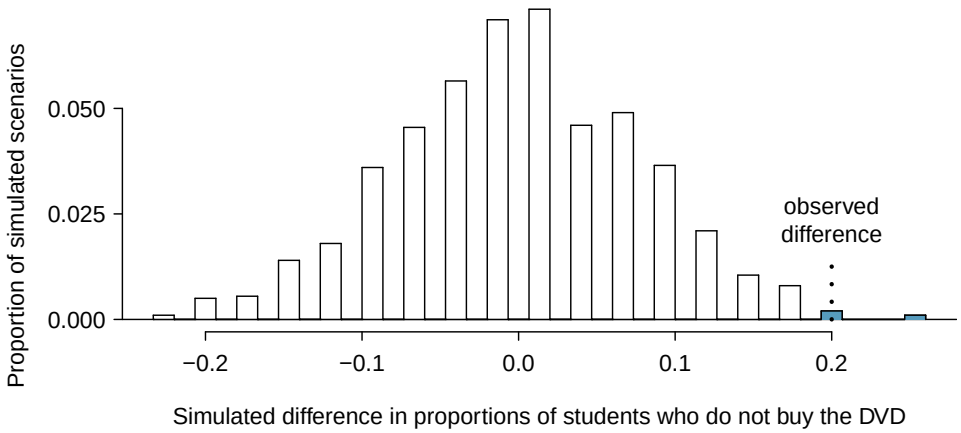


Figure 2.8: A histogram of 1,000 chance differences produced under the null hypothesis, H_0. Histograms like this one are a more convenient representation of data or results when there are a large number of observations.

2.3.1 Hypothesis testing in the US court system

Example 2.8 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 no reasonable doubt regarding the person's guilt. That is, the skeptical perspective (null hypothesis) is that the person is innocent until evidence is presented that convinces the jury that the person is guilty (alternative hypothesis).

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Notice that 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.

2.3.2 p-value and statistical significance

In Section 2.1 we encountered a study from the 1970's that explored whether there was strong evidence that women were less likely to be promoted than men. The research question – are females discriminated against in promotion decisions made by male managers? – was framed in the context of hypotheses:

H_0: Gender has no effect on promotion decisions.

H_A: Women are discriminated against in promotion decisions.

The null hypothesis (H_0) was a perspective of no difference. The data, summarized on page 62, provided a point estimate of a 29.2% difference in recommended promotion rates between men and women. We determined that such a difference from chance alone would be rare: it would only happen about 2 in 100 times. When results like these are inconsistent with H_0, we reject H_0 in favor of H_A. Here, we concluded there was discrimination against women.

The 2-in-100 chance is what we call a **p-value**, which is a probability quantifying the strength of the evidence against the null hypothesis and in favor of the alternative.

p-value

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, such as a difference in proportions, to help compute the p-value and evaluate the hypotheses. This summary value that is used to compute the p-value is often called the **test statistic**.

Example 2.9 In the gender discrimination study, the difference in discrimination rates was our test statistic. What was the test statistic in the opportunity cost study covered in Section 2.2?

The test statistic in the opportunity cost study was the difference in the proportion of students who decided against the DVD purchase in the treatment and control groups. In each of these examples, the point estimate of the difference in proportions was used as the test statistic.

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, called the **significance level** and often represented by α (the Greek letter *alpha*), is typically set to $\alpha = 0.05$, but can vary depending on the field or the application. Using a significance level of $\alpha = 0.05$ in the discrimination study, we can say that the data provided statistically significant evidence against the null hypothesis.

α
significance level of a hypothesis test

Statistical significance
We say that the data provide **statistically significant** evidence against the null hypothesis if the p-value is less than some reference value, usually $\alpha = 0.05$.

Example 2.10 In the opportunity cost study in Section 2.2, we analyzed an experiment where study participants were 20% less likely to continue with a DVD purchase if they were reminded that the money, if not spent on the DVD, could be used for other purchases in the future. 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 also a good idea to deviate from the standard. We'll discuss when to choose a threshold different than 0.05 in Section 2.3.4.

2.3.3 Decision errors

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 Table 2.9.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Table 2.9: Four different scenarios for hypothesis tests.

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. Since we rejected the null hypothesis in the gender discrimination and opportunity cost studies, it is possible that we made a Type 1 Error in one or both of those studies. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

Example 2.11 In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Table 2.9 may be useful.

If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. A Type 2 Error means the court failed to reject H_0 (i.e. failed to convict the person) when she was in fact guilty (H_A true).

Guided Practice 2.12 Consider the opportunity cost study where we concluded students were less likely to make a DVD purchase if they were reminded that money not spent now could be spent later. What would a Type 1 Error represent in this context?[9]

Example 2.13 How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?

To lower the Type 1 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 2 Errors.

[9] Making a Type 1 Error in this context would mean that reminding students that money not spent now can be spent later does not affect their buying habits, despite the strong evidence (the data suggesting otherwise) found in the experiment. 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.

⊙ **Guided Practice 2.14** How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[10]

The example and guided practice above provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

2.3.4 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is sometimes helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01 or 0.001). Under this scenario, we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring the alternative H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then 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 1 or Type 2 Error.

2.3.5 Introducing two-sided hypotheses

So far we have explored whether women were discriminated against and whether a simple trick could make students a little thriftier. In these two case studies, we've actually ignored some possibilities:

- What if *men* are actually discriminated against?
- What if the money trick actually makes students *spend more*?

These possibilities weren't considered in our hypotheses or analyses. This may have seemed natural since the data pointed in the directions in which we framed the problems. However, there are two dangers if we ignore possibilities that disagree with our data or that conflict with our worldview:

1. Framing an alternative hypothesis simply to match the direction that the data point will generally inflate the Type 1 Error rate. After all the work we've done (and will continue to do) to rigorously control the error rates in hypothesis tests, careless construction of the alternative hypotheses can disrupt that hard work. We'll explore this topic further in Section 2.3.6.
2. If we only use alternative hypotheses that agree with our worldview, then we're going to be subjecting ourselves to **confirmation bias**, which means we are looking for data that supports our ideas. That's not very scientific, and we can do better!

[10] To lower the Type 2 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 1 Error rate.

The previous hypotheses we've seen are called **one-sided hypothesis tests** because they only explored one direction of possibilities. Such hypotheses are appropriate when we are exclusively interested in the single direction, but usually we want to consider all possibilities. To do so, let's learn about **two-sided hypothesis tests** in the context of a new study that examines the impact of using blood thinners on patients who have undergone CPR.

Cardiopulmonary resuscitation (CPR) is a procedure used on individuals suffering a heart attack when other emergency resources are unavailable. This procedure is helpful in providing some blood circulation to keep a person alive, but CPR chest compressions can also cause internal injuries. Internal bleeding and other injuries that can result from CPR complicate additional treatment efforts. For instance, blood thinners may be used to help release a clot that is causing the heart attack once a patient arrives in the hospital. However, blood thinners negatively affect internal injuries.

Here we consider an experiment with patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.[11] Each patient was randomly assigned to either receive a blood thinner (treatment group) or not receive a blood thinner (control group). The outcome variable of interest was whether the patient survived for at least 24 hours.

Example 2.15 Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival rate of people who do not receive a blood thinner (corresponding to the control group) and p_t represent the survival rate for people receiving a blood thinner (corresponding to the treatment group).

We want to understand whether blood thinners are helpful or harmful. We'll consider both of these possibilities using a two-sided hypothesis test.

H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group. $p_t - p_c = 0$.

H_A: Blood thinners have an impact on survival, either positive or negative, but not zero. $p_t - p_c \neq 0$.

There were 50 patients in the experiment who did not receive a blood thinner and 40 patients who did. The study results are shown in Table 2.10.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Table 2.10: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

Guided Practice 2.16 What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$. [12]

[11] *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial*, by Böttiger et al., The Lancet, 2001.

[12] Observed control survival rate: $p_c = \frac{11}{50} = 0.22$. Treatment survival rate: $p_t = \frac{14}{40} = 0.35$. Observed difference: $\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13$.

According to the point estimate, for patients who have undergone CPR outside of the hospital, an additional 13% of these patients survive when they are treated with blood thinners. However, we wonder if this difference could be easily explainable by chance.

As we did in our past two studies this chapter, we will simulate what type of differences we might see from chance alone under the null hypothesis. By randomly assigning "simulated treatment" and "simulated control" stickers to the patients' files, we get a new grouping. If we repeat this simulation 10,000 times, we can build a **null distribution** of the differences shown in Figure 2.11.

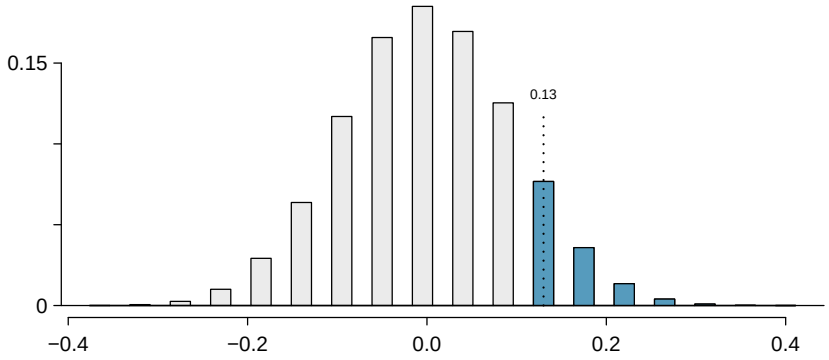


Figure 2.11: Null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The shaded right tail shows observations that are at least as large as the observed difference, 0.13.

The right tail area is about 0.13. (Note: it is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) However, contrary to how we calculated the p-value in previous studies, the p-value of this test is not 0.13!

The p-value is defined as the chance we observe a result at least as favorable to the alternative hypothesis as the result (i.e. the difference) we observe. In this case, any differences less than or equal to -0.13 would also provide equally strong evidence favoring the alternative hypothesis as a difference of 0.13. A difference of -0.13 would correspond to 13% higher survival rate in the control group than the treatment group. In Figure 2.12 we've also shaded these differences in the left tail of the distribution. These two shaded tails provide a visual representation of the p-value for a two-sided test.

For a two-sided test, take the single tail (in this case, 0.13) and double it to get the p-value: 0.26. Since this p-value is larger than 0.05, we do not reject the null hypothesis. That is, we do not find statistically significant evidence that the blood thinner has any influence on survival of patients who undergo CPR prior to arriving at the hospital.

Default to a two-sided test
We want to be rigorous and keep an open mind when we analyze data and evidence. Use a one-sided hypothesis test only if you truly have interest in only one direction.

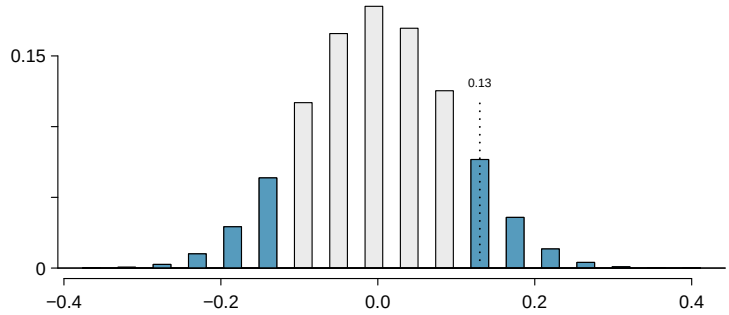


Figure 2.12: Null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. All values that are at least as extreme as +0.13 but in either direction away from 0 are shaded.

Computing a p-value for a two-sided test
First compute the p-value for one tail of the distribution, then double that value to get the two-sided p-value. That's it!

2.3.6 Controlling the Type 1 Error rate

It is never okay to change two-sided tests to one-sided tests after observing the data. We explore the consequences of ignoring this advice in the next example.

Example 2.17 Using $\alpha = 0.05$, we show that freely switching from two-sided tests to one-sided tests will lead us to make twice as many Type 1 Errors as intended.

Suppose we are interested in finding any difference from 0. We've created a smooth-looking **null distribution** representing differences due to chance in Figure 2.13.

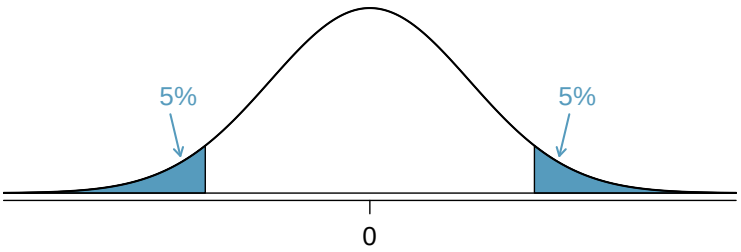


Figure 2.13: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example 2.17 when $\alpha = 0.05$.

Suppose the sample difference was larger than 0. Then if we can flip to a one-sided test, we would use H_A: difference > 0. Now if we obtain any observation in the upper 5% of the distribution, we would reject H_0 since the p-value would just be a the single tail. Thus, if the null hypothesis is true, we incorrectly reject the null hypothesis about 5% of the time when the sample mean is above the null value, as shown in Figure 2.13.

Suppose the sample difference was smaller than 0. Then if we change to a one-sided test, we would use H_A: difference < 0. If the observed difference falls in the lower 5% of the figure, we would reject H_0. That is, if the null hypothesis is true, then we would observe this situation about 5% of the time.

By examining these two scenarios, we can determine that we will make a Type 1 Error $5\% + 5\% = 10\%$ of the time if we are allowed to swap to the "best" one-sided test for the data. This is twice the error rate we prescribed with our significance level: $\alpha = 0.05$ (!).

Caution: Hypothesis tests should be set up *before* seeing the data
After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Hypotheses should be set up *before* observing the data.

2.3.7 How to use a hypothesis test

Frame the research question in terms of hypotheses. Hypothesis tests are appropriate for research questions that can be summarized in two competing hypotheses. The null hypothesis (H_0) usually represents a skeptical perspective or a perspective of no difference. The alternative hypothesis (H_A) usually represents a new view or a difference.

Collect data with an observational study or experiment. If a research question can be formed into two hypotheses, we can collect data to run a hypothesis test. If the research question focuses on associations between variables but does not concern causation, we would run an observational study. If the research question seeks a causal connection between two or more variables, then an experiment should be used.

Analyze the data. Choose an analysis technique appropriate for the data and identify the p-value. So far, we've only seen one analysis technique: randomization. Throughout the rest of this textbook, we'll encounter several new methods suitable for many other contexts.

Form a conclusion. Using the p-value from the analysis, determine whether the data provide statistically significant evidence against the null hypothesis. Also, be sure to write the conclusion in plain language so casual readers can understand the results.

2.4 Simulation case studies

Randomization is a statistical technique suitable for evaluating whether a difference in sample proportions is due to chance. In this section, we explore the situation where we focus on a single proportion, and we introduce a new simulation method.

2.4.1 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 average complication rate for liver donor surgeries in the US is about 10%, but her clients have had only 3 complications in the 62 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 2.18 We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.

The sample proportion for the complication rate is 3 complications divided by the 62 surgeries the consultant has worked on: $\hat{p} = 3/62 = 0.048$.

Example 2.19 Is it possible to assess the consultant's claim using the data?

No. The claim is 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.

While it is not possible to assess the causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Example 2.20 We're going to conduct a hypothesis test for this setting. Should the test be one-sided or two-sided?

The setting has been framed in the context of the consultant being helpful, but what if the consultant actually performed worse than the average? Would we care? More than ever! Since we care about a finding in either direction, we should run a two-sided test.

Guided Practice 2.21 Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.[13]

Parameter for a hypothesis test

A **parameter** for a hypothesis test is the "true" value of interest. We typically estimate the parameter using a point estimate from a sample of data.

For example, we estimate the probability p of a complication for a client of the medical consultant by examining the past complications rates of her clients:

$$\hat{p} = 3/62 = 0.048 \qquad \text{is used to estimate} \qquad p$$

Null value of a hypothesis test

The **null value** is the reference value for the parameter in H_0, and it is sometimes represented with the parameter's label with a subscript 0, e.g. p_0 (just like H_0).

[13] H_0: There is no association between the consultant's contributions and the clients' complication rate. That is, the complication rate for the consultant's clients is equal to the US average of 10%. In statistical language, $p = 0.10$. H_A: Patients who work with the consultant have a complication rate different than 10%, i.e. $p \neq 0.10$.

In the medical consultant case study, the parameter is p and the null value is $p_0 = 0.10$. We will use the p-value to quantify the possibility of a sample proportion ($\hat{p}$) this far from the null value. The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis were true. Just like we did using randomization for a difference in proportions, here we can simulate 62 new patients to see what result might happen if the complication rate was 0.10.

Each client can be simulated using 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 62 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.

Guided Practice 2.22 In a simulation of 62 patients, about how many would we expect to have had a complication?[14]

We conducted such a simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication: $\hat{p}_{sim} = 5/62 = 0.081$.

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 2.14 shows the null distribution from these 10,000 simulations. The simulated proportions that are less than or equal to $\hat{p} = 0.048$ are shaded. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$, which represents a fraction 0.1222 of our simulations:

$$\text{left tail} = \frac{\text{Number of observed simulations with } \hat{p}_{sim} \leq 0.048}{10000} = \frac{1222}{10000} = 0.1222$$

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:[15]

$$\text{p-value} = 2 \times \text{left tail} = 2 \times 0.1222 = 0.2444$$

Guided Practice 2.23 Because the p-value is 0.2444, 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.[16]

Example 2.24 Does the conclusion in Guided Practice 2.23 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.

No. It might be that the consultant's work is associated with a lower or higher risk of complications. However, the data did not provide enough information to reject the null hypothesis.

[14] About 10% of the patients (6.2 on average) in the simulation will have a complication, though we will see a little variation from one simulation to the next.

[15] This doubling approach is preferred even when the distribution isn't symmetric, as in this case.

[16] The data do not provide strong evidence that the consultant's work is associated with a lower or higher rate of surgery complications than the general rate of 10%.

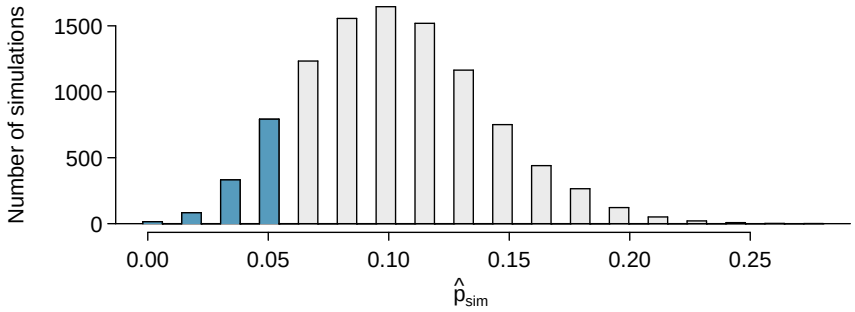


Figure 2.14: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail contains 12.22% of the simulations. We double this value to get the p-value.

2.4.2 Tappers and listeners

Here's a game you can try with your friends or family: pick a simple, well-known song, tap that tune on your desk, and see if the other person can guess the song. In this simple game, you are the tapper, and the other person is the listener.

A Stanford University graduate student named Elizabeth Newton conducted an experiment using the tapper-listener game.[17] In her study, she recruited 120 tappers and 120 listeners into the study. About 50% of the tappers expected that the listener would be able to guess the song. Newton wondered, is 50% a reasonable expectation?

Newton's research question can be framed into two hypotheses:

H_0: The tappers are correct, and generally 50% of the time listeners are able to guess the tune. $p = 0.50$

H_A: The tappers are incorrect, and either more than or less than 50% of listeners will be able to guess the tune. $p \neq 0.50$

In Newton's study, only 3 out of 120 listeners ($\hat{p} = 0.025$) were able to guess the tune! From the perspective of the null hypothesis, we might wonder, how likely is it that we would get this result from chance alone? That is, what's the chance we would happen to see such a small fraction if H_0 were true and the true correct-guess rate is 0.50?

We will again use a simulation. To simulate 120 games under the null hypothesis where $p = 0.50$, we could flip a coin 120 times. Each time the coin came up heads, this could represent the listener guessing correctly, and tails would represent the listener guessing incorrectly. For example, we can simulate 5 tapper-listener pairs by flipping a coin 5 times:

H	H	T	H	T
Correct	Correct	Wrong	Correct	Wrong

After flipping the coin 120 times, we got 56 heads for $\hat{p}_{sim} = 0.467$. As we did with the randomization technique, seeing what would happen with one simulation isn't enough. In order to evaluate whether our originally observed proportion of 0.025 is unusual or not, we should generate more simulations. Here we've repeated this simulation ten times:

0.558 0.517 0.467 0.458 0.525 0.425 0.458 0.492 0.550 0.483

[17] This case study is described in *Made to Stick* by Chip and Dan Heath. Little known fact: the teaching principles behind many OpenIntro resources are based on *Made to Stick*.

As before, we'll run a total of 10,000 simulations using a computer. Figure 2.15 shows the results of these simulations. Even in these 10,000 simulations, we don't see any results close to 0.025.

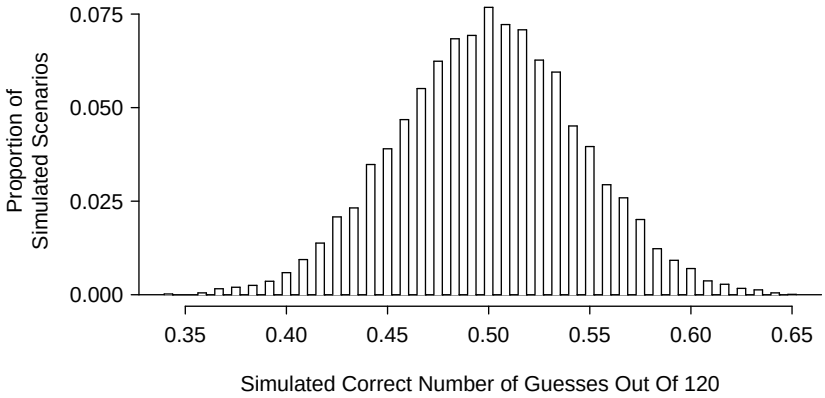


Figure 2.15: Results from 10,000 simulations of the tapper-listener study where guesses are correct half of the time.

⊙ **Guided Practice 2.25** What is the p-value for the hypothesis test?[18]

⊙ **Guided Practice 2.26** Do the data provide statistically significant evidence against the null hypothesis? State an appropriate conclusion in the context of the research question.[19]

2.5 Central Limit Theorem

We've encountered four case studies so far this chapter. While they differ in the settings, in their outcomes, and also in the technique we've used to analyze the data, they all have something in common: the general shape of the null distribution.

2.5.1 Null distribution from the case studies

Figure 2.16 shows the null distributions in each of the four case studies where we ran 10,000 simulations. In the case of the opportunity cost study, which originally had just 1,000 simulations, we've included an additional 9,000 simulations.

⊙ **Guided Practice 2.27** Describe the shape of the distributions and note anything that you find interesting.[20]

[18] The p-value is the chance of seeing the data summary or something more in favor of the alternative hypothesis. Since we didn't observe anything even close to just 3 correct, the p-value will be small, around 1-in-10,000 or smaller.

[19] The p-value is less than 0.05, so we reject the null hypothesis. There is statistically significant evidence, and the data provide strong evidence that the chance a listener will guess the correct tune is less than 50%.

[20] In general, the distributions are reasonably symmetric. The case study for the medical consultant is the only distribution with any evident skew.

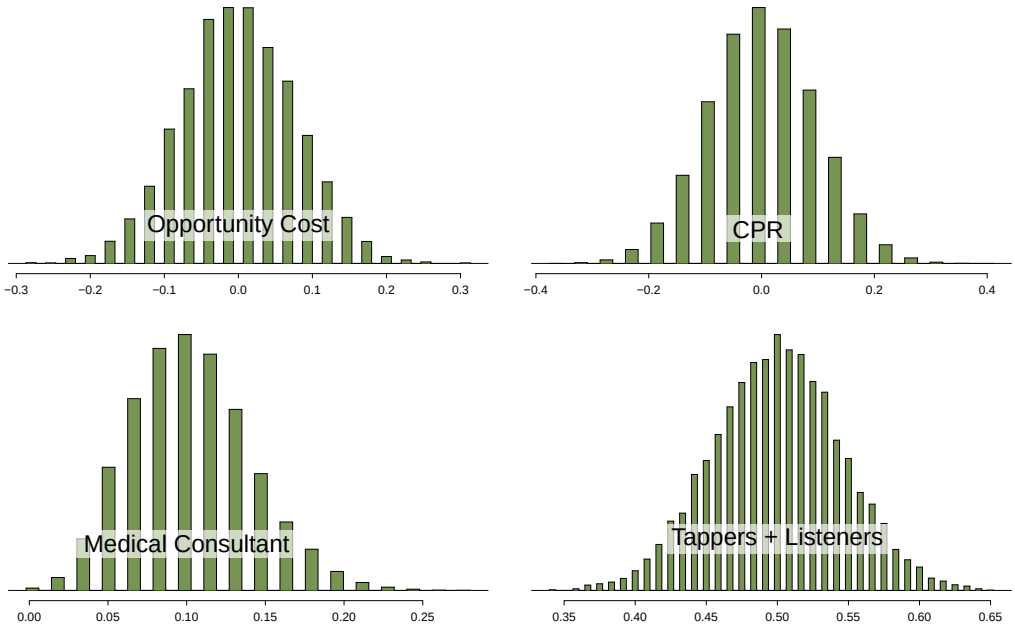


Figure 2.16: The null distribution for each of the four case studies presented in Sections 2.2-2.4.

As we observed in Chapter 1, it's common for distributions to be skewed or contain outliers. However, the null distributions we've so far encountered have all looked somewhat similar and, for the most part, symmetric. They all resemble a bell-shaped curve. This is not a coincidence, but rather, is guaranteed by mathematical theory.

Central Limit Theorem for proportions
If we look at a proportion (or difference in proportions) and the scenario satisfies certain conditions, then the sample proportion (or difference in proportions) will appear to follow a bell-shaped curve called the *normal distribution*.

An example of a perfect normal distribution is shown in Figure 2.17. Imagine laying a normal curve over each of the four null distributions in Figure 2.16. While the mean and standard deviation may change for each plot, the general shape remains roughly intact.

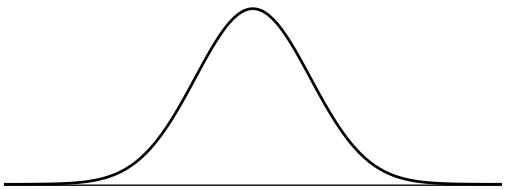


Figure 2.17: A normal curve.

Mathematical theory guarantees that a sample proportion or a difference in sample proportions will follow something that resembles a normal distribution when certain conditions are met. These conditions fall into two categories:

Observations in the sample are independent. Independence is guaranteed when we take a random sample from a population. It can also be guaranteed if we randomly divide individuals into treatment and control groups.

The sample is large enough. The sample size cannot be too small. What qualifies as "small" differs from one context to the next, and we'll provide suitable guidelines for proportions in Chapter 3.

So far we've had no need for the normal distribution. We've been able to answer our questions somewhat easily using simulation techniques. However, soon this will change. Simulating data can be non-trivial. For example, some scenarios that we will encounter in Chapters 5 and 6 would require complex simulations. Instead, the normal distribution and other distributions like it offer a general framework that applies to a very large number of settings.

2.5.2 Examples of future settings we will consider

Below we introduce three new settings where the normal distribution will be useful but constructing suitable simulations can be difficult.

Example 2.28 The opportunity cost study determined that students are thriftier if they are reminded that saving money now means they can spend the money later. The study's point estimate for the estimated impact was 20%, meaning 20% fewer students would move forward with a DVD purchase in the study scenario. However, as we've learned, point estimates aren't perfect – they only provide an approximation of the truth.

It would be useful if we could provide a *range of plausible values* for the impact, more formally known as a **confidence interval**. It is often difficult to construct a reliable confidence interval in many situations using simulations.[21] However, doing so is reasonably straightforward using the normal distribution. We'll tackle this topic in Section 2.8.

Example 2.29 Book prices were collected for 73 courses at UCLA in Spring 2010. Data were collected from both the UCLA Bookstore and Amazon. The differences in these prices are shown in Figure 2.18. The mean difference in the price of the books was $12.76, and we might wonder, does this provide strong evidence that the prices differ between the two book sellers?

Here again we can apply the normal distribution, this time in the context of numerical data. We'll explore this example and construct such a hypothesis test in Section 4.2.

[21] The percentile bootstrap method has been put forward as an alternative. However, simulations show that this method is consistently less robust than than the normal distribution. For more information, visit openintro.org/bootstrap.

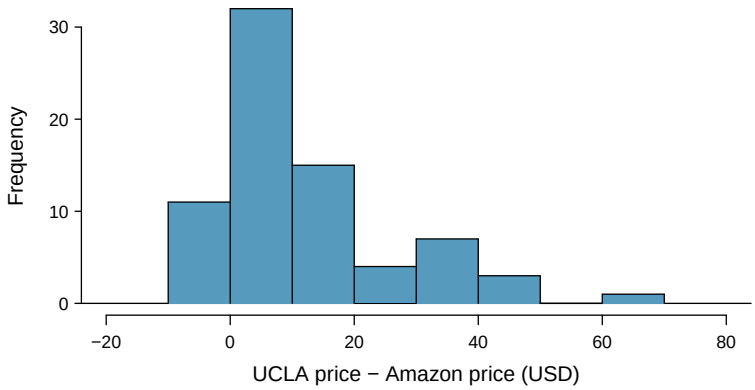


Figure 2.18: Histogram of the difference in price for each book sampled. These data are strongly skewed.

Example 2.30 Elmhurst College in Illinois released anonymized data for family income and financial support provided by the school for Elmhurst's first-year students in 2011. Figure 2.19 shows a *regression line* fit to a scatterplot of a sample of the data. One question we will ask is, do the data show a real trend, or is the trend we observe reasonably explained by chance?

In Chapter 5 we'll learn how to apply least squares regression to quantify the trend and quantify whether or not that trend can be explained by chance alone. For this case study, we could again use the normal distribution to help us answer this question.

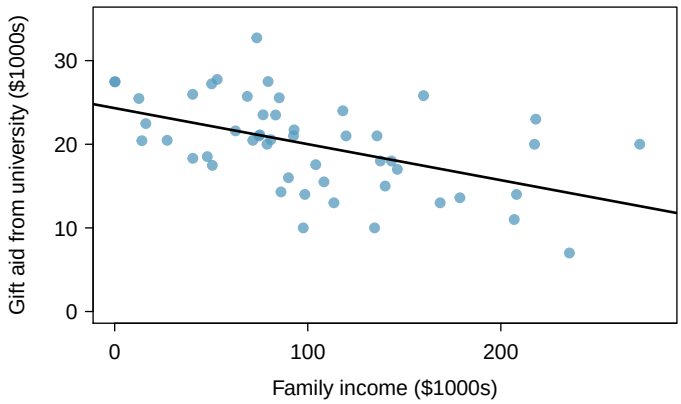


Figure 2.19: Gift aid and family income for a random sample of 50 first-year students from Elmhurst College, shown with a regression line.

These examples highlight the value of the normal distribution approach. However, before we can apply the normal distribution to statistical inference, it is necessary to become familiar with the mechanics of the normal distribution. In Section 2.6 we discuss characteristics of the normal distribution, explore examples of data that follow a normal distribution, and learn a new plotting technique that is useful for evaluating whether a data set roughly follows the normal distribution. In Sections 2.7 and 2.8, we apply this new knowledge in the context of hypothesis tests and confidence intervals.

2.6 Normal distribution

Among all the distributions we see in statistics, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. It is so common that people often know it as the **normal curve**, **normal model**, or **normal distribution**.[22] Under certain conditions, sample proportions, sample means, and differences can be modeled using the normal distribution. Additionally, some variables such as SAT scores and heights of US adult males closely follow the normal distribution.

Normal distribution facts

Many summary statistics and variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. We will use it in data exploration and to solve important problems in statistics.

In this section, we will discuss the normal distribution in the context of data to (1) become familiar with normal distribution techniques and (2) learn how to evaluate whether data are nearly normal. In Sections 2.7-2.8 and beyond, we'll move our discussion to focus on applying the normal distribution and other related distributions to model point estimates for hypothesis tests and for constructing confidence intervals.

2.6.1 Normal distribution model

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 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 2.20 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 2.21 shows these distributions on the same axis.

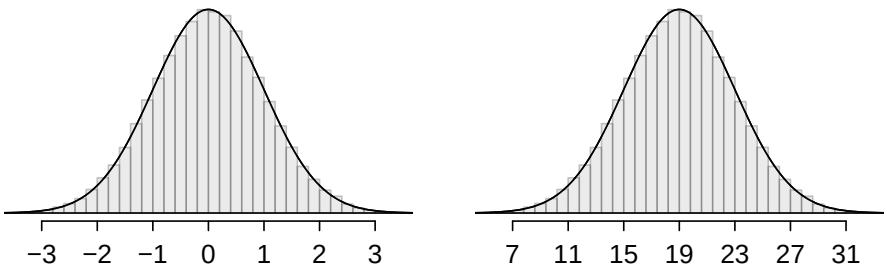


Figure 2.20: Both curves represent the normal distribution, however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

[22] It is also introduced as the Gaussian distribution after Frederic Gauss, the first person to formalize its mathematical expression.

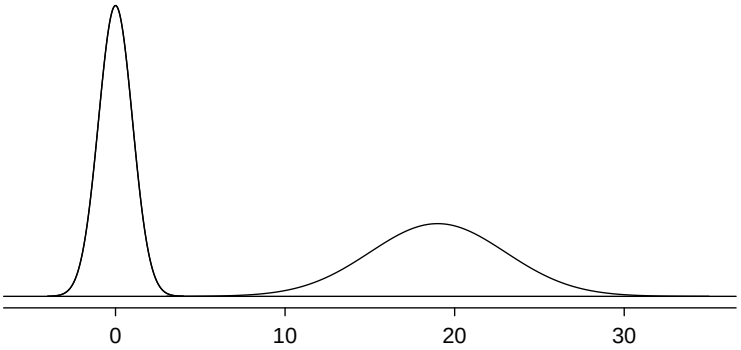


Figure 2.21: The normal models shown in Figure 2.20 but plotted together and on the same scale.

$N(\mu, \sigma)$
Normal dist. with mean μ & st. dev. σ

If a normal distribution has mean μ and standard deviation σ, we may write the distribution as $N(\mu, \sigma)$. The two distributions in Figure 2.21 can be written as

$$N(\mu = 0, \sigma = 1) \quad \text{and} \quad N(\mu = 19, \sigma = 4)$$

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**.

⊙ **Guided Practice 2.31** Write down the short-hand for a normal distribution with (a) mean 5 and standard deviation 3, (b) mean -100 and standard deviation 10, and (c) mean 2 and standard deviation 9. [23]

2.6.2 Standardizing with Z scores

● **Example 2.32** Table 2.22 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 1800 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: $1500 + 300 = 1800$. Tom is 0.6 standard deviations above the mean on the ACT: $21 + 0.6 \times 5 = 24$. In Figure 2.23, 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	1500	21
SD	300	5

Table 2.22: Mean and standard deviation for the SAT and ACT.

Example 2.32 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.

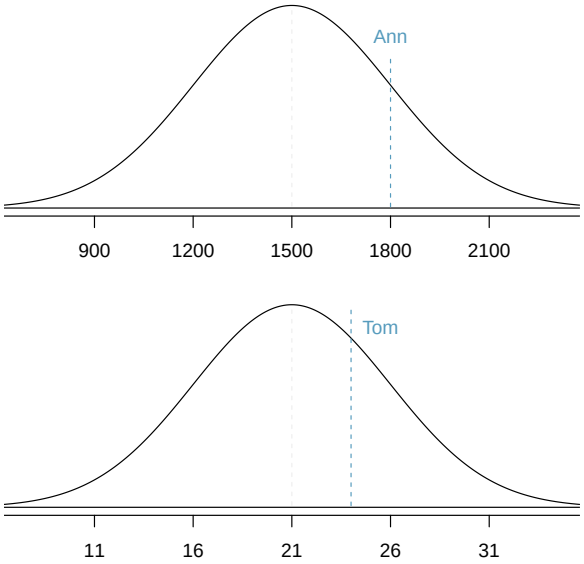


Figure 2.23: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

Z
Z score, the standardized observation

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 $N(\mu, \sigma)$, we define the Z score mathematically as

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

Using $\mu_{SAT} = 1500$, $\sigma_{SAT} = 300$, and $x_{Ann} = 1800$, we find Ann's Z score:

$$Z_{Ann} = \frac{x_{Ann} - \mu_{SAT}}{\sigma_{SAT}} = \frac{1800 - 1500}{300} = 1$$

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

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

⊙ **Guided Practice 2.33** Use Tom's ACT score, 24, along with the ACT mean and standard deviation to compute his Z score.[24]

Observations above the mean always have positive Z scores while those below the mean have negative Z scores. If an observation is equal to the mean (e.g. SAT score of 1500), then the Z score is 0.

[23] (a) $N(\mu = 5, \sigma = 3)$. (b) $N(\mu = -100, \sigma = 10)$. (c) $N(\mu = 2, \sigma = 9)$.

[24] $Z_{Tom} = \frac{x_{Tom} - \mu_{ACT}}{\sigma_{ACT}} = \frac{24-21}{5} = 0.6$

⊙ **Guided Practice 2.34** Let X represent a random variable from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$. (a) Find the Z score of x. (b) Use the Z score to determine how many standard deviations above or below the mean x falls.[25]

⊙ **Guided Practice 2.35** Head lengths of brushtail possums follow a nearly normal distribution with mean 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.[26]

We can use Z scores to roughly identify which observations are more unusual than others. One 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 2.36** Which of the observations in Guided Practice 2.35 is more unusual?[27]

2.6.3 Normal probability table

● **Example 2.37** Ann from Example 2.32 earned a score of 1800 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 2.24. 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 2.24: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

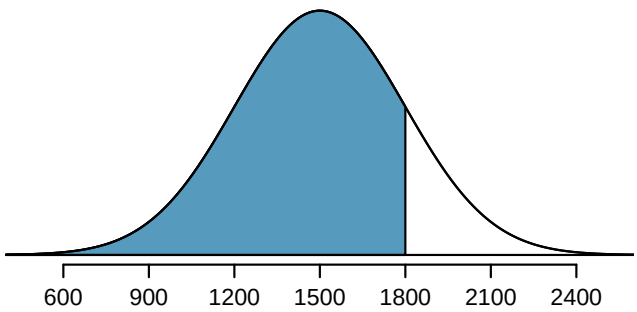


Figure 2.24: The normal model for SAT scores, shading the area of those individuals who scored below Ann.

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.

[25] (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.

[26] For $x_1 = 95.4$ mm: $Z_1 = \frac{x_1-\mu}{\sigma} = \frac{95.4-92.6}{3.6} = 0.78$. For $x_2 = 85.8$ mm: $Z_2 = \frac{85.8-92.6}{3.6} = -1.89$.

[27] Because the *absolute value* of Z score for the second observation is larger than that of the first, the second observation has a more unusual head length.

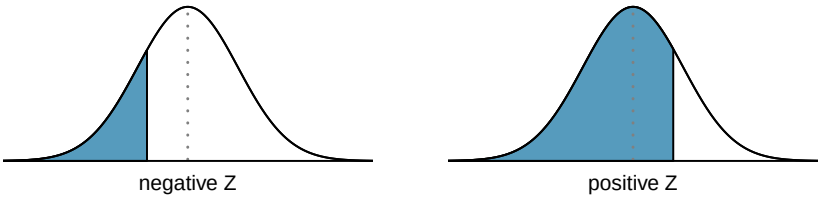


Figure 2.25: The area to the left of Z represents the percentile of the observation.

	Second decimal place of Z									
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.26: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.43$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

A normal probability table is given in Appendix C.1 on page 339 and abbreviated in Table 2.26. 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 Table 2.26: 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.

We can also find the Z score associated with a percentile. For example, to identify Z for the 80^{th} 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 80^{th} percentile by combining the row and column Z values: 0.84.

⊙ **Guided Practice 2.38** Determine the proportion of SAT test takers who scored better than Ann on the SAT.[28]

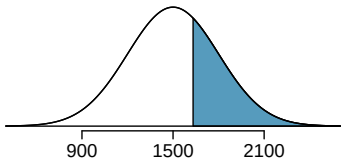
[28] If 84% had lower scores than Ann, the number of people who had better scores must be 16%. (Generally ties are ignored when the normal model, or any other continuous distribution, is used.)

2.6.4 Normal probability examples

Cumulative SAT scores are approximated well by a normal model, $N(\mu = 1500, \sigma = 300)$.

Example 2.39 Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1630 on her SATs?

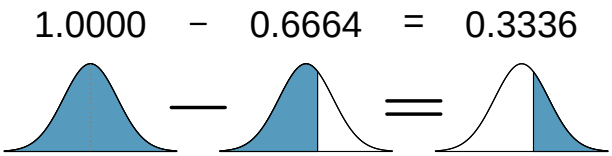
First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the chance she scores above 1630, so we shade this upper tail:



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 $\mu = 1500$, $\sigma = 300$, and the cutoff value $x = 1630$, the Z score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1630 - 1500}{300} = \frac{130}{300} = 0.43$$

We look up the percentile of $Z = 0.43$ in the normal probability table shown in Table 2.26 or in Appendix C.1 on page 339, which yields 0.6664. However, the percentile describes those who had a Z score *lower* than 0.43. To find the area *above* $Z = 0.43$, we compute one minus the area of the lower tail:



The probability Shannon scores at least 1630 on the SAT is 0.3336.

TIP: 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.

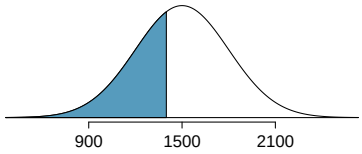
After drawing a figure to represent the situation, identify the Z score for the observation of interest.

Guided Practice 2.40 If the probability of Shannon scoring at least 1630 is 0.3336, then what is the probability she scores less than 1630? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[29]

[29] We found the probability in Example 2.39: 0.6664. A picture for this exercise is represented by the shaded area below "0.6664" in Example 2.39.

Example 2.41 Edward earned a 1400 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 1400. These are the scores to the left of 1400.



Identifying the mean $\mu = 1500$, the standard deviation $\sigma = 300$, and the cutoff for the tail area $x = 1400$ makes it easy to compute the Z score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1400 - 1500}{300} = -0.33$$

Using the normal probability table, identify the row of -0.3 and column of 0.03, which corresponds to the probability 0.3707. Edward is at the 37^{th} percentile.

Guided Practice 2.42 Use the results of Example 2.41 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[30]

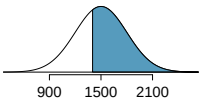
TIP: areas to the right
The normal probability table in most books gives the area to the left. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

Guided Practice 2.43 Stuart earned an SAT score of 2100. Draw a picture for each part. (a) What is his percentile? (b) What percent of SAT takers did better than Stuart?[31]

Based on a sample of 100 men,[32] the heights of male adults between the ages 20 and 62 in the US is nearly normal with mean 70.0" and standard deviation 3.3".

Guided Practice 2.44 Mike is 5'7" and Jim is 6'4". (a) What is Mike's height percentile? (b) What is Jim's height percentile? Also draw one picture for each part.[33]

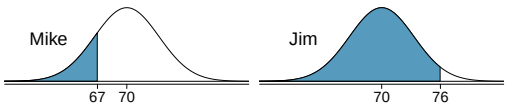
[30] If Edward did better than 37% of SAT takers, then about 63% must have done better than him.



[31] Numerical answers: (a) 0.9772. (b) 0.0228.

[32] This sample was taken from the USDA Food Commodity Intake Database.

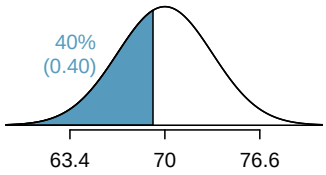
[33] First put the heights into inches: 67 and 76 inches. Figures are shown below. (a) $Z_{Mike} = \frac{67-70}{3.3} = -0.91 \rightarrow 0.1814$. (b) $Z_{Jim} = \frac{76-70}{3.3} = 1.82 \rightarrow 0.9656$.



The last several problems have focused on finding the probability or percentile for a particular observation. What if you would like to know the observation corresponding to a particular percentile?

Example 2.45 Erik's height is at the 40^{th} percentile. How tall is he?

As always, first draw the picture.



In this case, the lower tail probability is known (0.40), which can be shaded on the diagram. We want to find the observation that corresponds to this value. As a first step in this direction, we determine the Z score associated with the 40^{th} percentile.

Because the percentile is below 50%, we know Z will be negative. Looking in the negative part of the normal probability table, we search for the probability *inside* the table closest to 0.4000. We find that 0.4000 falls in row -0.2 and between columns 0.05 and 0.06. Since it falls closer to 0.05, we take this one: $Z = -0.25$.

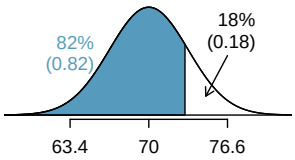
Knowing $Z_{Erik} = -0.25$ and the population parameters $\mu = 70$ and $\sigma = 3.3$ inches, the Z score formula can be set up to determine Erik's unknown height, labeled x_{Erik}:

$$-0.25 = Z_{Erik} = \frac{x_{Erik} - \mu}{\sigma} = \frac{x_{Erik} - 70}{3.3}$$

Solving for x_{Erik} yields the height 69.18 inches. That is, Erik is about 5'9" (this is notation for 5-feet, 9-inches).

Example 2.46 What is the adult male height at the 82^{nd} percentile?

Again, we draw the figure first.



Next, we want to find the Z score at the 82^{nd} percentile, which will be a positive value. Looking in the Z table, we find Z falls in row 0.9 and the nearest column is 0.02, i.e. $Z = 0.92$. Finally, the height x is found using the Z score formula with the known mean μ, standard deviation σ, and Z score $Z = 0.92$:

$$0.92 = Z = \frac{x - \mu}{\sigma} = \frac{x - 70}{3.3}$$

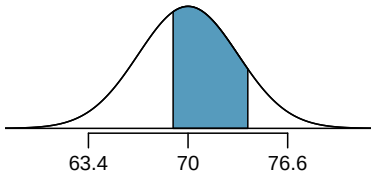
This yields 73.04 inches or about 6'1" as the height at the 82^{nd} percentile.

⊙ **Guided Practice 2.47** (a) What is the 95^{th} percentile for SAT scores? (b) What is the 97.5^{th} percentile of the male heights? As always with normal probability problems, first draw a picture.[34]

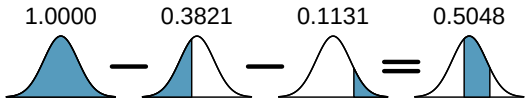
⊙ **Guided Practice 2.48** (a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)? (b) What is the probability that a male adult is shorter than 5'9" (69 inches)?[35]

● **Example 2.49** What is the probability that a random adult male is between 5'9" and 6'2"?

These heights correspond to 69 inches and 74 inches. First, draw the figure. The area of interest is no longer an upper or lower tail.



The total area under the curve is 1. If we find the area of the two tails that are not shaded (from Guided Practice 2.48, these areas are 0.3821 and 0.1131), then we can find the middle area:



That is, the probability of being between 5'9" and 6'2" is 0.5048.

⊙ **Guided Practice 2.50** What percent of SAT takers get between 1500 and 2000?[36]

⊙ **Guided Practice 2.51** What percent of adult males are between 5'5" and 5'7"?[37]

2.6.5 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.

[34] Remember: draw a picture first, then find the Z score. (We leave the pictures to you.) The Z score can be found by using the percentiles and the normal probability table. (a) We look for 0.95 in the probability portion (middle part) of the normal probability table, which leads us to row 1.6 and (about) column 0.05, i.e. $Z_{95} = 1.65$. Knowing $Z_{95} = 1.65$, $\mu = 1500$, and $\sigma = 300$, we setup the Z score formula: $1.65 = \frac{x_{95} - 1500}{300}$. We solve for x_{95}: $x_{95} = 1995$. (b) Similarly, we find $Z_{97.5} = 1.96$, again setup the Z score formula for the heights, and calculate $x_{97.5} = 76.5$.

[35] Numerical answers: (a) 0.1131. (b) 0.3821.

[36] This is an abbreviated solution. (Be sure to draw a figure!) First find the percent who get below 1500 and the percent that get above 2000: $Z_{1500} = 0.00 \rightarrow 0.5000$ (area below), $Z_{2000} = 1.67 \rightarrow 0.0475$ (area above). Final answer: $1.0000 - 0.5000 - 0.0475 = 0.4525$.

[37] 5'5" is 65 inches. 5'7" is 67 inches. Numerical solution: $1.000 - 0.0649 - 0.8183 = 0.1168$, i.e. 11.68%.

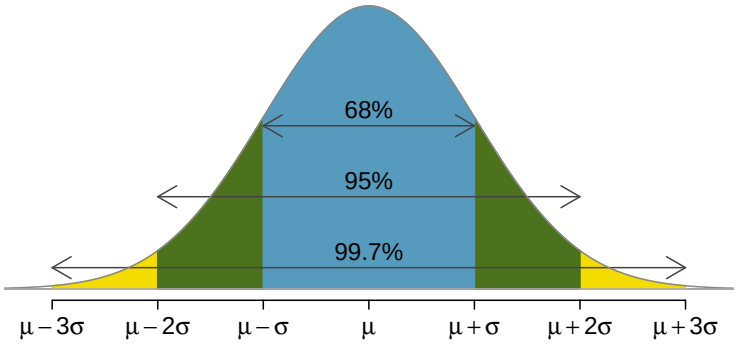


Figure 2.27: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

⊙ **Guided Practice 2.52** 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 $Z = 2$.[38]

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-30,000. For 5 and 6 standard deviations, it is about 1-in-3.5 million and 1-in-1 billion, respectively.

⊙ **Guided Practice 2.53** SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$. (a) About what percent of test takers score 900 to 2100? (b) What percent score between 1500 and 2100?[39]

2.6.6 Evaluating the normal approximation

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males. While using a normal model can be extremely convenient and helpful, 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.

Example 2.45 suggests 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, which 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 2.28. The sample

[38] 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 $Z = 3$.

[39] (a) 900 and 2100 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 900 and 2100. (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 900 to 1500 while 47.5% score between 1500 and 2100.

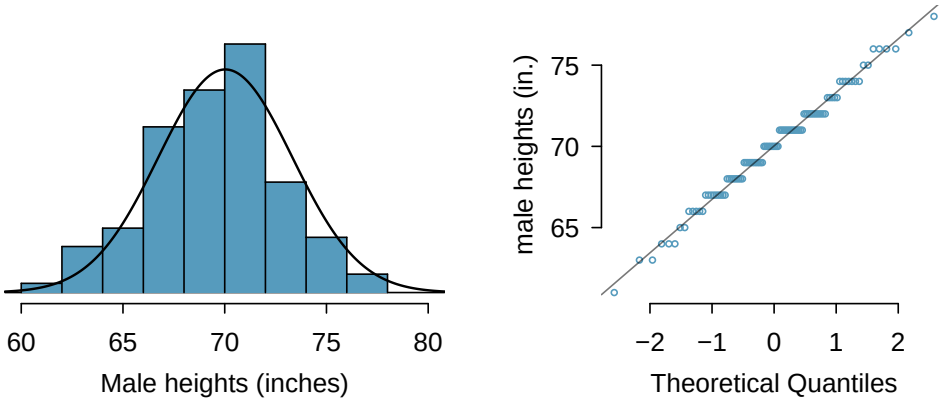


Figure 2.28: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

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**.[40], shown in the right panel of Figure 2.28. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.

Example 2.54 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 2.29. 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.

[40] Also commonly called a **quantile-quantile plot**.

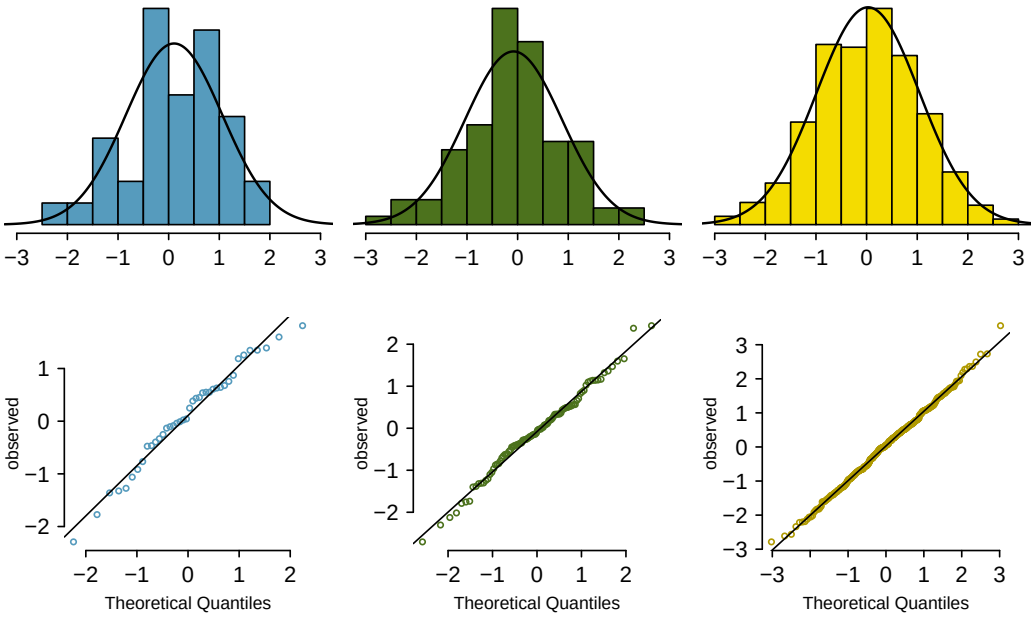


Figure 2.29: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

Example 2.55 Are NBA player heights normally distributed? Consider all 435 NBA players from the 2008-9 season presented in Figure 2.30. [41]

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 2.54 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 2.56 Can we approximate poker winnings by a normal distribution? We consider the poker winnings of an individual over 50 days. A histogram and normal probability plot of these data are shown in Figure 2.31.

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 2.54, it is apparent that these data show very strong deviations from the normal model.

[41] These data were collected from http://www.nba.com.

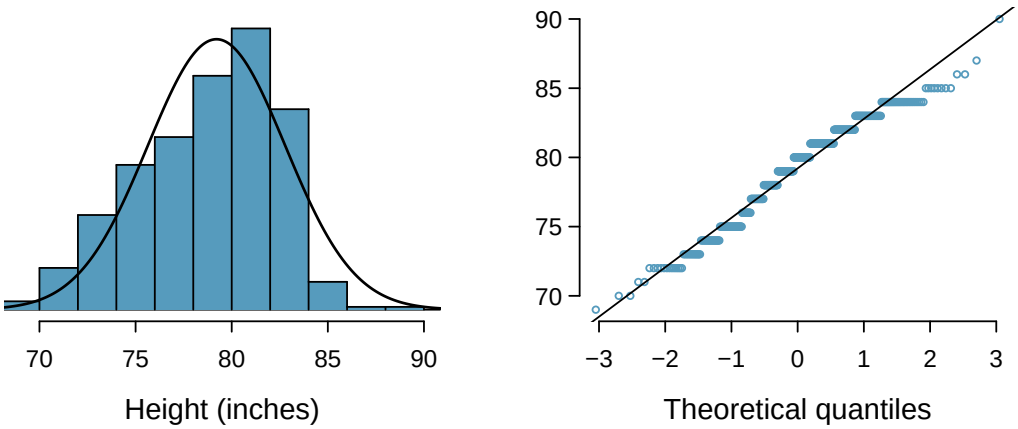


Figure 2.30: Histogram and normal probability plot for the NBA heights from the 2008-9 season.

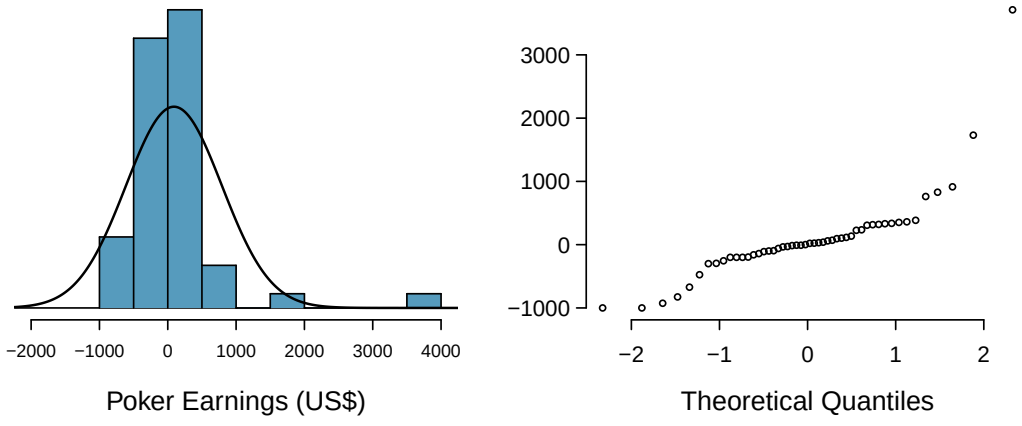


Figure 2.31: A histogram of poker data with the best fitting normal plot and a normal probability plot.

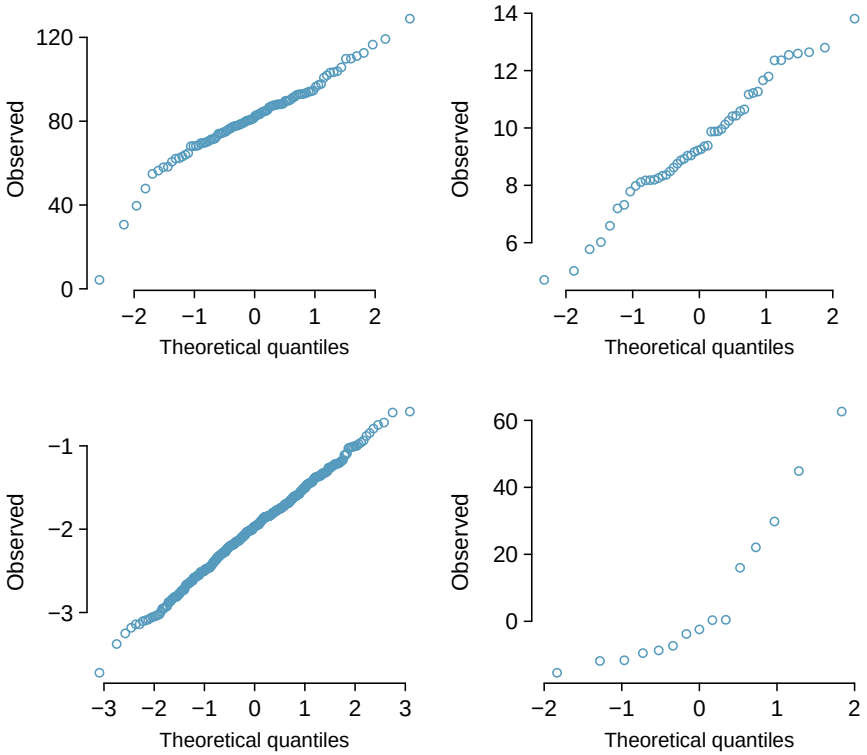


Figure 2.32: Four normal probability plots for Guided Practice 2.57.

⊙ **Guided Practice 2.57** Determine which data sets represented in Figure 2.32 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.[42]

⊙ **Guided Practice 2.58** Figure 2.33 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?[43]

[42] 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.

[43] 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.

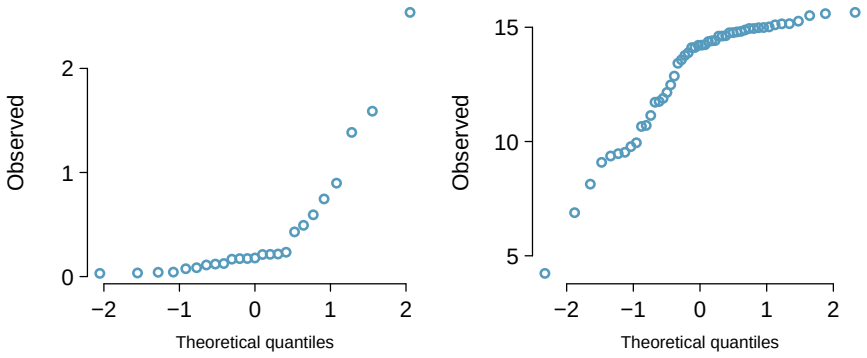


Figure 2.33: Normal probability plots for Guided Practice 2.58.

2.7 Applying the normal model

The approach for using the normal model in the context of inference is very similar to the practice of applying the model to individual observations that are nearly normal. We will replace null distributions we previously obtained using the randomization or simulation techniques and verify the results once again using the normal model. When the sample size is sufficiently large, this approximation generally provides us with the same conclusions.

2.7.1 Standard error

Point estimates vary from sample to sample, and we quantify this variability with what is called the **standard error (SE)**. The standard error is equal to the standard deviation associated with the estimate. So, for example, if we used the standard deviation to quantify the variability of a point estimate from one sample to the next, this standard deviation would be called the standard error of the point estimate.

The way we determine the standard error varies from one situation to the next. However, typically it is determined using a formula based on the Central Limit Theorem.

2.7.2 Normal model application: opportunity cost

In Section 2.2 we were introduced to the opportunity cost study, which found that students became thriftier when they were reminded that not spending money now means the money can be spent on other things in the future. Let's re-analyze the data in the context of the normal distribution and compare the results.

Figure 2.34 summarizes the null distribution as determined using the randomization method. The best fitting normal distribution for the null distribution has a mean of 0. We can calculate the standard error of this distribution by borrowing a formula that we will become familiar with in Section 3.2, but for now let's just take the value $SE = 0.078$ as a given. Recall that the point estimate of the difference was 0.20, as shown in the plot. Next, we'll use the normal distribution approach to compute the two-tailed p-value.

As we learned in Section 2.6, it is helpful to draw and shade a picture of the normal distribution so we know precisely what we want to calculate. Here we want to find the area of the two tails representing the p-value.

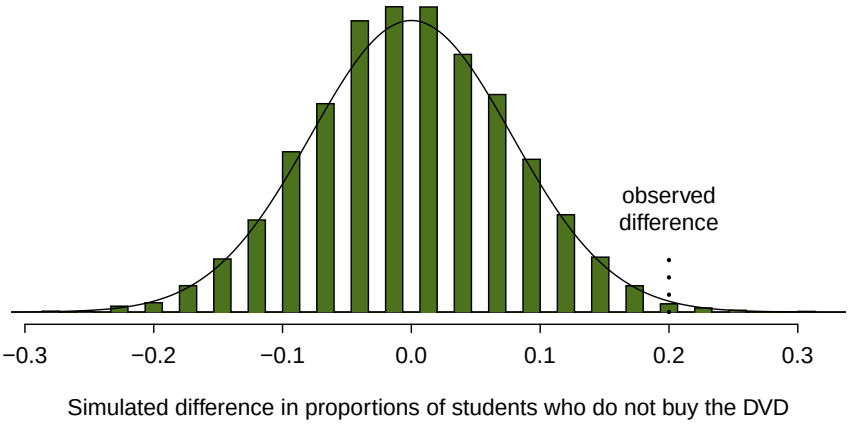
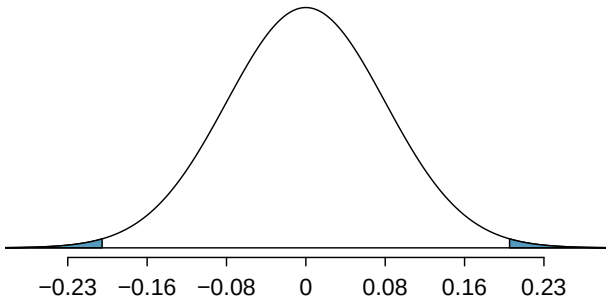


Figure 2.34: Null distribution of differences with an overlaid normal curve for the opportunity cost study. 10,000 simulations were run for this figure.



Next, we can calculate the Z score using the observed difference, 0.20, and the two model parameters. The standard error, $SE = 0.078$, is the equivalent of the model's standard deviation.

$$Z = \frac{\text{observed difference} - 0}{SE} = \frac{0.20 - 0}{0.078} = 2.56$$

We can either look up $Z = 2.56$ in the normal probability table or use statistical software to determine the right tail area: 0.0052, which is about the same as what we got for the right tail using the randomization approach (0.0065). Doubling this value yields the total area in the two tails and the p-value for the hypothesis test: 0.01. As before, since the p-value is less than 0.05, we conclude that the treatment did indeed impact students' spending.

Z score in a hypothesis test

In the context of a hypothesis test, the Z score for a point estimate is

$$Z = \frac{\text{point estimate} - \text{null value}}{SE}$$

The standard error in this case is the equivalent of the standard deviation of the point estimate, and the null value comes from the null hypothesis.

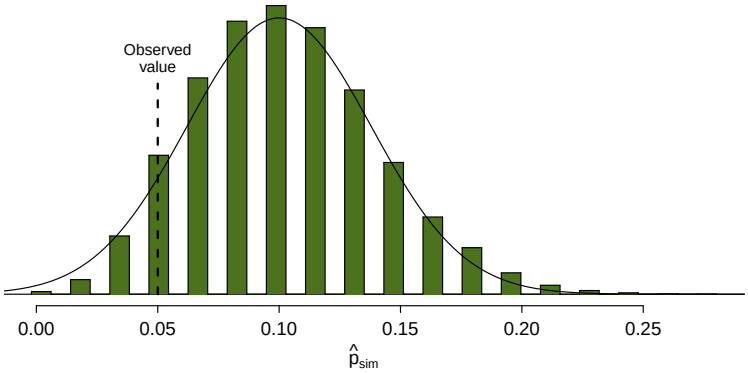


Figure 2.35: The null distribution for $\hat{p}$, created from 10,000 simulated studies, along with the best-fitting normal model.

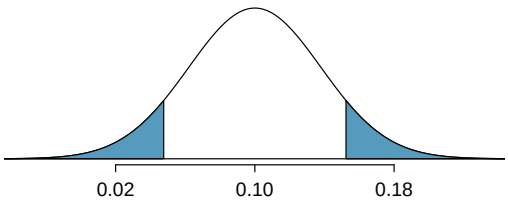
We have confirmed that the randomization approach we used earlier and the normal distribution approach provide almost identical p-values and conclusions in the opportunity cost case study. Next, let's turn our attention to the medical consultant case study.

2.7.3 Normal model application: medical consultant

In Section 2.4 we learned about a medical consultant who reported that only 3 of her 62 clients who underwent a liver transplant had complications, which is less than the more common complication rate of 0.10. As in the other case studies, we identified a suitable null distribution using a simulation approach, as shown in Figure 2.35. Here we have added the best-fitting normal curve to the figure, which has a mean of 0.10. Borrowing a formula that we'll encounter in Chapter 3, the standard error of this distribution was also computed: $SE = 0.038$.

In the previous analysis, we obtained a p-value of 0.2444, and we will try to reproduce that p-value using the normal distribution approach. However, before we begin, we want to point out a simple detail that is easy to overlook: the null distribution we earlier generated is slightly skewed, and the distribution isn't that smooth. In fact, the normal distribution only sort-of fits this model. We'll discuss this discrepancy more in a moment.

We'll again begin by creating a picture. Here a normal distribution centered at 0.10 with a standard error of 0.038.



Next, we can calculate the Z score using the observed complication rate, $\hat{p} = 0.048$ along with the mean and standard deviation of the normal model. Here again, we use the standard error for the standard deviation.

$$Z = \frac{\hat{p} - p_0}{SE_{\hat{p}}} = \frac{0.048 - 0.10}{0.038} = -1.37$$

Identifying $Z = -1.37$ in the normal probability table or using statistical software, we can determine that the left tail area is 0.0853. Doubling this value yields the total area in the two tails: about 0.17. This is the estimated p-value for the hypothesis test. However, there's a problem: this is very different than the earlier p-value we computed: 0.2444.

The discrepancy is explained by normal model's poor representation of the null distribution in Figure 2.35. As noted earlier, the null distribution from the simulations is not very smooth, and the distribution itself is slightly skewed. That's the bad news. The good news is that we can foresee these problems using some simple checks. We'll learn about these checks in the following chapters.

In Section 2.5 we noted that the two common requirements to apply the Central Limit Theorem are (1) the observations in the sample must be independent, and (2) the sample must be sufficiently large. The guidelines for this particular situation – which we will learn in Section 3.1 – would have alerted us that the normal model was a poor approximation.

2.7.4 Conditions for applying the normal model

The success story in this section was the application of the normal model in the context of the opportunity cost data. However, the biggest lesson comes from our failed attempt to use the normal approximation in the medical consultant case study.

Statistical techniques are like a carpenter's tools. When used responsibly, they can produce amazing and precise results. However, if the tools are applied irresponsibly or under inappropriate conditions, they will produce unreliable results. For this reason, with every statistical method that we introduce in future chapters, we will carefully outline conditions when the method can reasonably be used. These conditions should be checked in each application of the technique.

2.8 Confidence intervals

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. In addition to supplying a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* for the parameter.

2.8.1 Capturing the population parameter

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.

⊙ **Guided Practice 2.59** If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[44]

[44] If we want to be more certain we will capture the fish, we might use a wider net. Likewise, we use a wider confidence interval if we want to be more certain that we capture the parameter.

2.8.2 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, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

Constructing a 95% confidence interval

When the sampling distribution of a point estimate can reasonably be modeled as normal, the point estimate we observe will be within 1.96 standard errors of the true value of interest about 95% of the time. Thus, a **95% confidence interval** for such a point estimate can be constructed:

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{2.60}$$

We can be **95% confident** this interval captures the true value.

⊙ **Guided Practice 2.61** Compute the area between -1.96 and 1.96 for a normal distribution with mean 0 and standard deviation 1. [45]

● **Example 2.62** The point estimate from the opportunity cost study was that 20% fewer students would buy a DVD if they were reminded that money not spent now could be spent later on something else. The point estimate from this study can reasonably be modeled with a normal distribution, and a proper standard error for this point estimate is $SE = 0.078$. Construct a 95% confidence interval.[46]

Since the conditions for the normal approximation have already been verified, we can move forward with the construction of the 95% confidence interval:

$$\text{point estimate} \ \pm \ 1.96 \times SE \quad \rightarrow \quad 0.20 \ \pm \ 1.96 \times 0.078 \quad \rightarrow \quad (0.047, 0.353)$$

We are 95% confident that the DVD purchase rate resulting from the treatment is between 4.7% and 35.3% lower than in the control group. Since this confidence interval does not contain 0, it is consistent with our earlier result where we rejected the notion of "no difference" using a hypothesis test.

In Section 1.1 we encountered an experiment that examined whether implanting a stent in the brain of a patient at risk for a stroke helps reduce the risk of a stroke. The results from the first 30 days of this study, which included 451 patients, are summarized in Table 2.36. These results are surprising! The point estimate suggests that patients who received stents may have a *higher* risk of stroke: $p_{trmt} - p_{ctrl} = 0.090$.

---

[45] We will leave it to you to draw a picture. The Z scores are $Z_{left} = -1.96$ and $Z_{right} = 1.96$. The area between these two Z scores is $0.9750 - 0.0250 = 0.9500$. This is where "1.96" comes from in the 95% confidence interval formula.

[46] We've used $SE = 0.078$ from the last section. However, it would more generally be appropriate to recompute the SE slightly differently for this confidence interval using the technique introduced in Section 3.2.1. Don't worry about this detail for now since the two resulting standard errors are, in this case, almost identical.

	stroke	no event	Total
treatment	33	191	224
control	13	214	227
Total	46	405	451

Table 2.36: Descriptive statistics for 30-day results for the stent study.

Example 2.63 Consider the stent study and results. The conditions necessary to ensure the point estimate $p_{trmt} - p_{ctrl} = 0.090$ is nearly normal have been verified for you, and the estimate's standard error is $SE = 0.028$. Construct a 95% confidence interval for the change in 30-day stroke rates from usage of the stent.

The conditions for applying the normal model have already been verified, so we can proceed to the construction of the confidence interval:

$$\text{point estimate} \pm 1.96 \times SE \quad \rightarrow \quad 0.090 \pm 1.96 \times 0.028 \quad \rightarrow \quad (0.035, 0.145)$$

We are 95% confident that implanting a stent in a stroke patient's brain increased the risk of stroke within 30 days by a rate of 0.035 to 0.145. This confidence interval can also be used in a way analogous to a hypothesis test: since the interval does not contain 0, it means the data provide statistically significant evidence that the stent used in the study *increases* the risk of stroke, contrary to what researchers had expected before this study was published!

As with hypothesis tests, confidence intervals are imperfect. About 1-in-20 properly constructed 95% confidence intervals will fail to capture the parameter of interest. Figure 2.37 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.

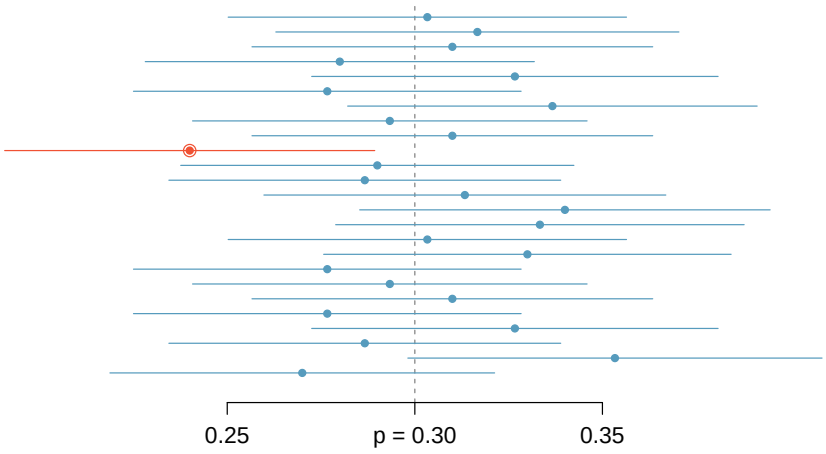


Figure 2.37: Twenty-five samples of size $n = 300$ were simulated when $p = 0.30$. For each sample, a confidence interval was created to try to capture the true proportion p. However, 1 of these 25 intervals did not capture $p = 0.30$.

⊙ **Guided Practice 2.64** In Figure 2.37, one interval does not contain the true proportion, $p = 0.3$. Does this imply that there was a problem with the simulations run?[47]

2.8.3 Changing the confidence level

Suppose we want to consider confidence intervals where the confidence level is somewhat higher than 95%: perhaps we would like a confidence level of 99%. Think back to the analogy about trying to catch a fish: if we want to be more sure that we will catch the fish, we should use a wider net. To create a 99% confidence level, we must also widen our 95% interval. On the other hand, if we want an interval with lower confidence, such as 90%, we could make our original 95% interval slightly slimmer.

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:

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{2.65}$$

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 data since the estimate is within 1.96 standard errors of the true value about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

⊙ **Guided Practice 2.66** If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[48]

To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58. Guided Practice 2.66 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of its mean. This approach – using the Z scores in the normal model to compute confidence levels – is appropriate when the point estimate is associated with a normal distribution and we can properly compute the standard error. Thus, the formula for a 99% confidence interval is

$$\text{point estimate} \ \pm \ 2.58 \times SE \tag{2.67}$$

The normal approximation is crucial to the precision of these confidence intervals. The next two chapters provides detailed discussions about when the normal model can safely be applied to a variety of situations. When the normal model is not a good fit, we will use alternative distributions that better characterize the sampling distribution.

[47] 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.

[48] 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 2.38.) To determine this probability, look up -2.58 and 2.58 in the normal probability table (0.0049 and 0.9951). Thus, there is a $0.9951 - 0.0049 \approx 0.99$ probability that the unobserved random variable X will be within 2.58 standard deviations of the mean.

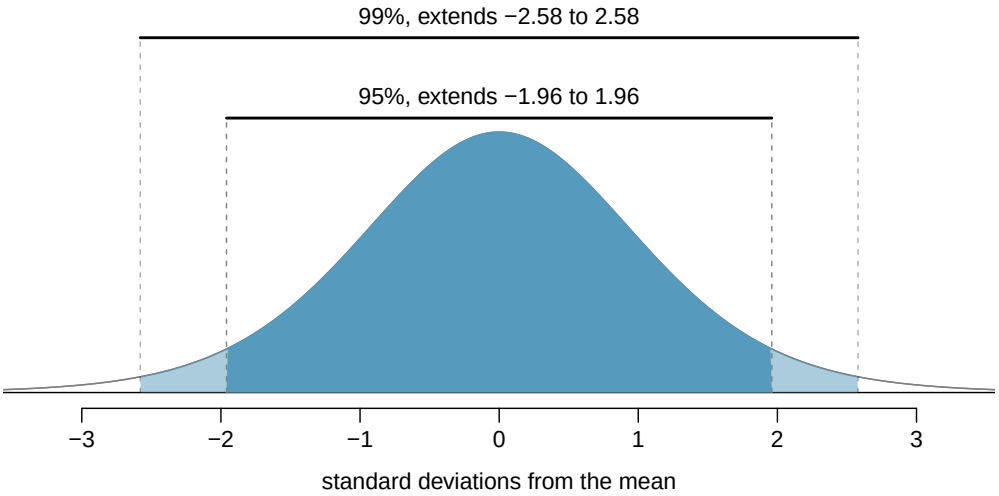


Figure 2.38: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

⊙ **Guided Practice 2.68** Create a 99% confidence interval for the impact of the stent on the risk of stroke using the data from Example 2.63. The point estimate is 0.090, and the standard error is $SE = 0.028$. It has been verified for you that the point estimate can reasonably be modeled by a normal distribution.[49]

Confidence interval for any confidence level
If the point estimate follows the normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z^\star \times SE$$

where $z^\star$ corresponds to the confidence level selected.

Figure 2.38 provides a picture of how to identify $z^\star$ based on a 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.

Margin of error
In a confidence interval, $z^\star \times SE$ is called the **margin of error**.

[49] Since the necessary conditions for applying the normal model have already been checked for us, we can go straight to the construction of the confidence interval: point estimate $\pm$ $2.58 \times SE \rightarrow (0.018, 0.162)$. We are 99% confident that implanting a stent in the brain of a patient who is at risk of stroke increases the risk of stroke within 30 days by a rate of 0.018 to 0.162 (assuming the patients are representative of the population).

⊙ **Guided Practice 2.69** In Example 2.63 we found that implanting a stent in the brain of a patient at risk for a stroke *increased* the risk of a stroke. The study estimated a 9% increase in the number of patients who had a stroke, and the standard error of this estimate was about $SE = 2.8\%$. Compute a 90% confidence interval for the effect.[50]

2.8.4 Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

> We are XX% confident that the population parameter is between...

Incorrect language might try to describe the confidence interval as capturing the population parameter with a certain probability. This is one of the most common errors: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval.

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 about capturing point estimates. Confidence intervals only attempt to capture population parameters.

[50] We must find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal model, $N(\mu = 0, \sigma = 1)$. We can look up -$z^\star$ in the normal probability table by looking for a lower tail of 5% (the other 5% is in the upper tail), thus $z^\star = 1.65$. The 90% confidence interval can then be computed as point estimate $\pm\ 1.65 \times SE \rightarrow (4.4\%, 13.6\%)$. (Note: the conditions for normality had earlier been confirmed for us.) That is, we are 90% confident that implanting a stent in a stroke patient's brain increased the risk of stroke within 30 days by 4.4% to 13.6%.

2.9 Exercises

2.9.1 Randomization case study: gender discrimination

2.1 Side effects of Avandia, Part I. 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.[51]

		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

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.

(a) 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.

(b) 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.

(c) The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

(d) 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.

[51] 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.

2.2 Heart transplants, Part II. Exercise 1.50 introduces the Stanford Heart Transplant Study. Of the 34 patients in the control group, 4 were alive at the end of the study. Of the 69 patients in the treatment group, 24 were alive. The contingency table below summarizes these results.

		Group		
		Control	Treatment	Total
Outcome	Alive	4	24	28
	Dead	30	45	75
	Total	34	69	103

(a) What proportion of patients in the treatment group and what proportion of patients in the control group died?

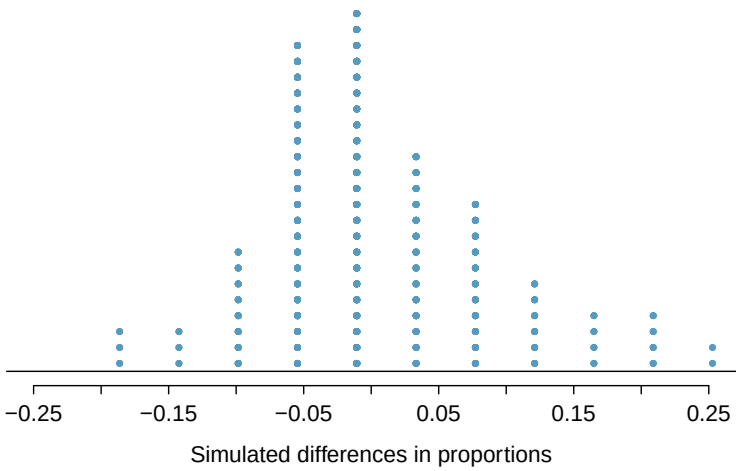
(b) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested? Use the same null and alternative hypothesis notation used in the section.

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 many 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?



2.9.2 Randomization case study: opportunity cost

2.3 Side effects of Avandia, Part II. Exercise 2.1 introduces a study that compares the rates of serious cardiovascular problems for diabetic patients on rosiglitazone and pioglitazone treatments. The table below summarizes the results of the study.

		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) What proportion of all patients had cardiovascular problems?

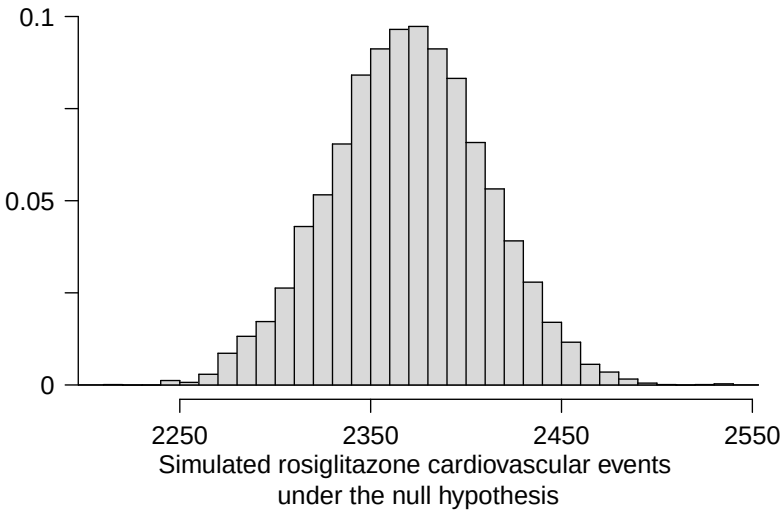
(b) If the type of treatment and having cardiovascular problems were independent (null hypothesis), about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

(c) 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 null hypothesis, 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 10,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Below is a relative frequency histogram of these counts.

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?



2.4 Sinusitis and antibiotics, Part II. Researchers studying the effect of antibiotic treatment compared to symptomatic treatment for acute sinusitis randomly assigned 166 adults diagnosed with sinusitis into two groups (as discussed in Exercise 1.2). Participants in the antibiotic group received a 10-day course of an antibiotic, and the rest received symptomatic treatments as a placebo. These pills had the same taste and packaging as the antibiotic. At the end of the 10-day period patients were asked if they experienced improvement in symptoms since the beginning of the study. The distribution of responses is summarized below.[52]

		Self-reported improvement in symptoms		
		Yes	No	Total
Treatment	Antibiotic	66	19	85
	Placebo	65	16	81
	Total	131	35	166

(a) What type of a study is this?

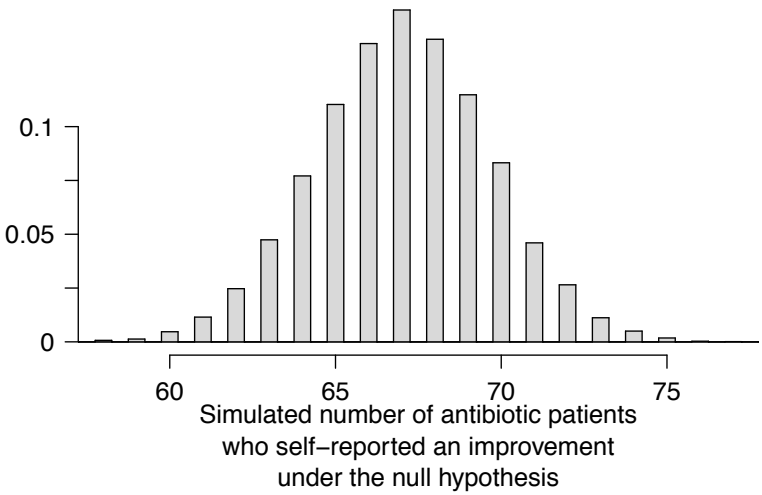
(b) Does this study make use of blinding?

(c) Compute the difference in the proportions of patients who self-reported an improvement in symptoms in the two groups: $\hat{p}_{antibiotic} - \hat{p}_{placebo}$.

(d) At first glance, does antibiotic or placebo appear to be more effective for the treatment of sinusitis? Explain your reasoning using appropriate statistics.

(e) There are two competing claims that this study is used to compare: the null hypothesis that the antibiotic has no impact and the alternative hypothesis that it has an impact. Write out these competing claims in easy-to-understand language and in the context of the application.

(f) Below is a histogram of simulation results computed under the null hypothesis. In each simulation, the summary value reported was the number of patients who received antibiotics and self-reported an improvement in symptoms. Write a conclusion for the hypothesis test in plain language. (Hint: Does the value observed in the study, 66, seem unusual in this distribution generated under the null hypothesis?)



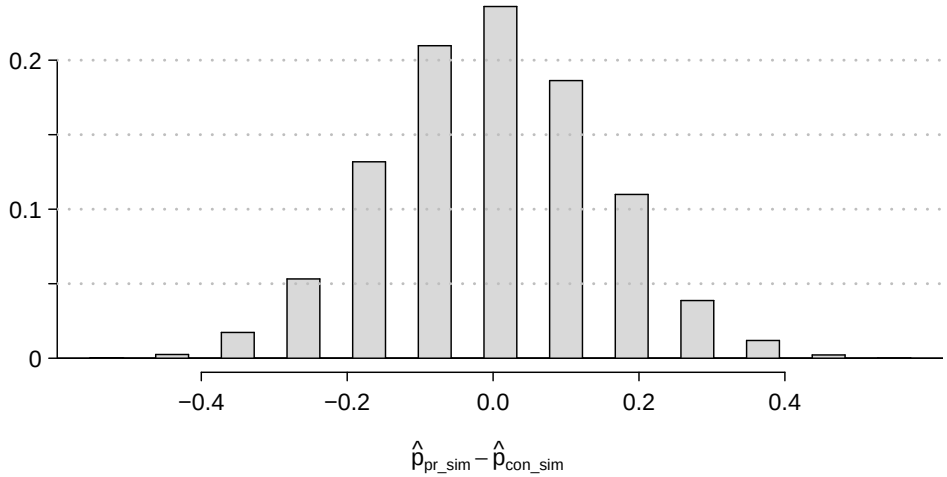
[52] J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

2.9.3 Hypothesis testing

2.5 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		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	No	15	10	25
	Total	20	25	45

A simulation was conducted to test if people react differently under the two scenarios. 10,000 simulated differences were generated to construct the null distribution shown. The value $\hat{p}_{pr,sim}$ represents the proportion of diners who intervened in the simulation for the provocatively dressed woman, and $\hat{p}_{con,sim}$ is the proportion for the conservatively dressed woman.



(a) What are the hypotheses? For the purposes of this exercise, you may assume that each observed person at the restaurant behaved independently, though we would want to evaluate this assumption more rigorously if we were reporting these results.

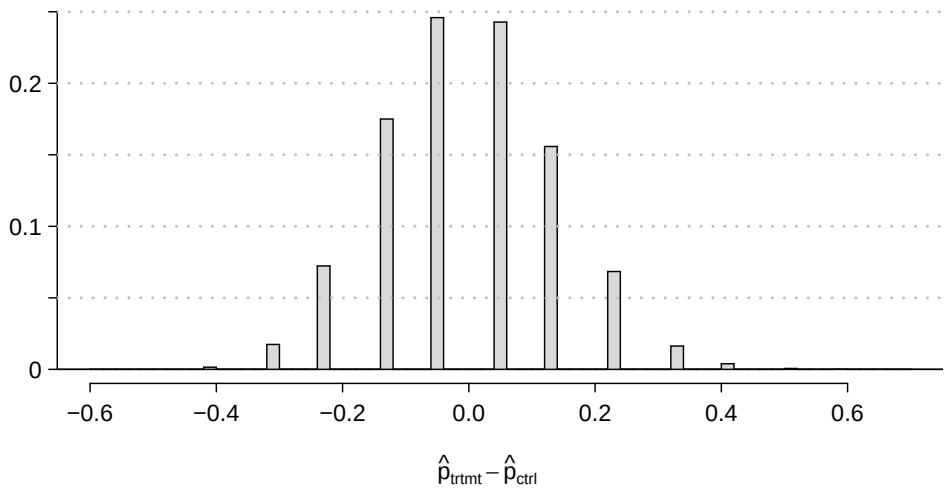
(b) Calculate the observed difference between the rates of intervention under the provocative and conservative scenarios: $\hat{p}_{pr} - \hat{p}_{con}$.

(c) Estimate the p-value using the figure above and determine the conclusion of the hypothesis test.

2.6 Is yawning contagious, Part I. An experiment conducted by the *MythBusters*, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.[53]

		Group		
		Treatment	Control	Total
Result	Yawn	10	4	14
	Not Yawn	24	12	36
	Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no influence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning as $\hat{p}_{trtmt,sim} - \hat{p}_{ctrl,sim}$. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.



(a) What are the hypotheses?

(b) Calculate the observed difference between the yawning rates under the two scenarios.

(c) Estimate the p-value using the figure above and determine the conclusion of the hypothesis test.

[53] MythBusters, Season 3, Episode 28.

2.7 Social experiment, Part II. In Exercise 2.5, we encountered a scenario where researchers were evaluating the impact of the way someone is dressed against the actions of people around them. In that exercise, researchers may have believed that dressing provocatively may reduce the chance of bystander intervention. One might be tempted to use a one-sided hypothesis test for this study. Discuss the drawbacks of doing so in 1-3 sentences.

2.8 Is yawning contagious, Part II. Exercise 2.6 describes an experiment by Myth Busters, where they examined whether a person yawning would affect whether others to yawn. The traditional belief is that yawning is contagious – one yawn can lead to another yawn, which might lead to another, and so on. In that exercise, there was the option of selecting a one-sided or two-sided test. Which would you recommend (or which did you choose)? Justify your answer in 1-3 sentences.

2.9.4 Simulation case studies

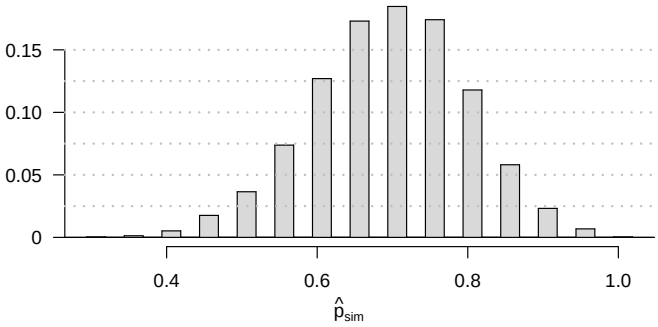
2.9 The Egyptian Revolution. A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the first couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.[54]

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

(c) Describe how to perform a simulation and, once you had results, how to estimate the p-value.

(d) Below is a histogram showing the distribution of $\hat{p}_{sim}$ in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.



2.10 Assisted Reproduction. Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g. in vitro fertilization). A 2008 report by the Centers for Disease Control and Prevention estimated that ART has been successful in leading to a live birth in 31% of cases[55]. A new fertility clinic claims that their success rate is higher than average. A random sample of 30 of their patients yielded a success rate of 40%. A consumer watchdog group would like to determine if this provides strong evidence to support the company's claim.

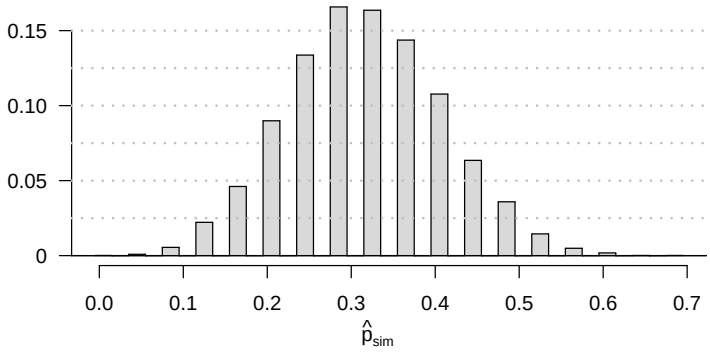
(a) Write the hypotheses to test if the success rate for ART at this clinic is significantly higher than the success rate reported by the CDC.

[54] Gallup Politics, Americans' Views of Egypt Sharply More Negative, data collected February 2-5, 2011.
[55] CDC. *2008 Assisted Reproductive Technology Report*.

(b) Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

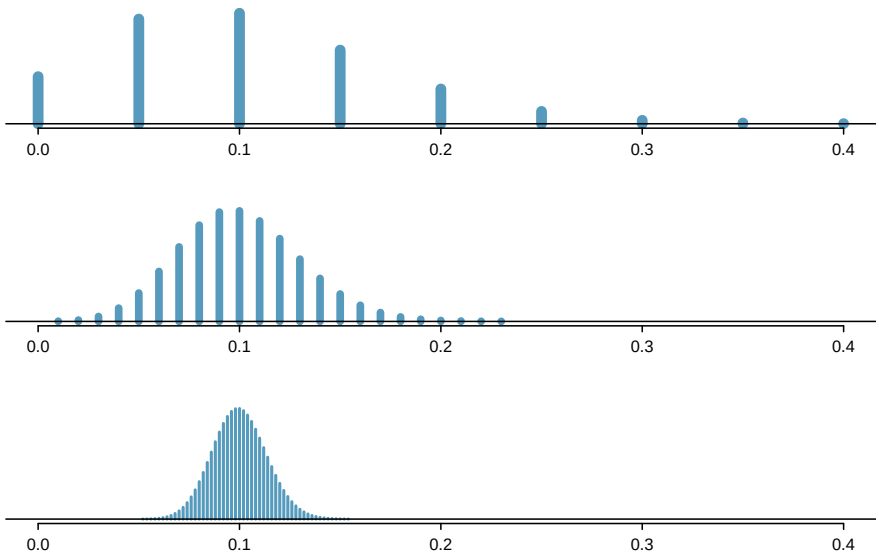
(c) Below is a histogram showing the distribution of $\hat{p}_{sim}$ in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

(d) After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.

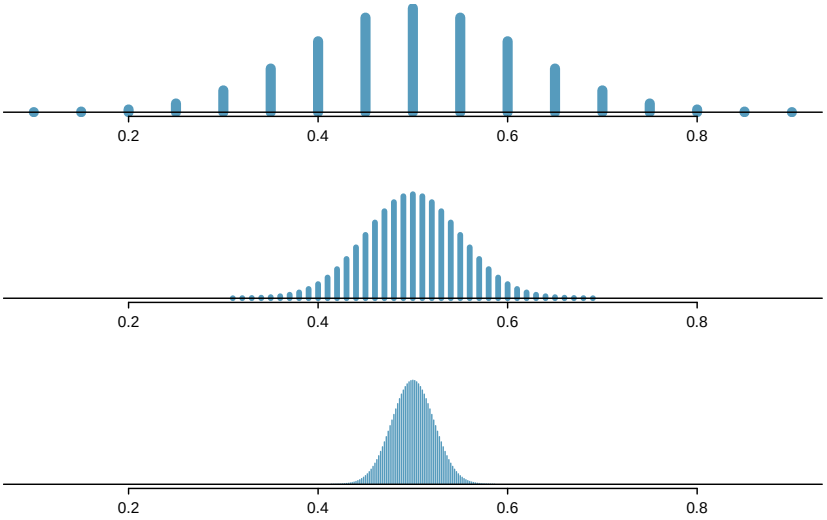


2.9.5 Central Limit Theorem

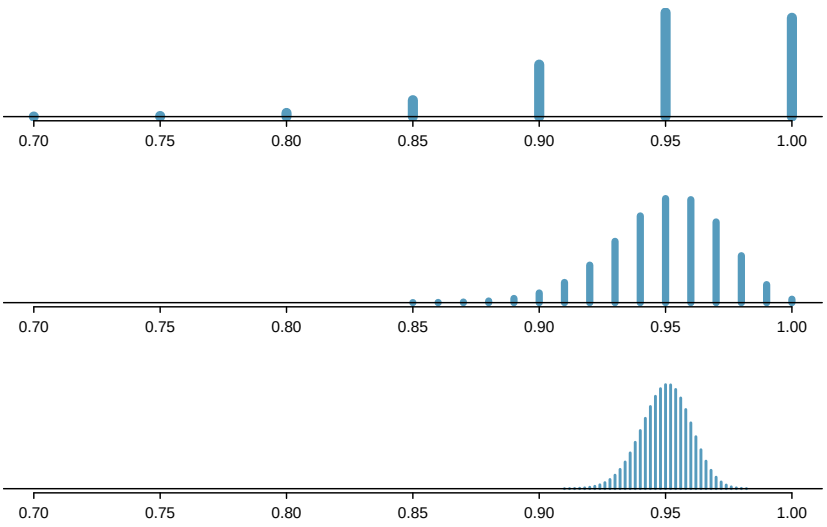
2.11 Distribution of $\hat{p}$. Suppose the true population proportion were $p = 0.1$. The figure below shows what the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$. What does each point (observation) in each of the samples represent? Describe how the distribution of the sample proportion, $\hat{p}$, changes as n becomes larger.



2.12 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.



2.13 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 $n = 500$. 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.



2.14 Re-examining the distributions of past exercises. Examine the distributions shown in Exercises 2.3, 2.4, 2.5, 2.6, 2.9, and 2.10. Which distributions look symmetric and bell-shaped? Which appear to be overly "discrete" (not very smooth)?

2.9.6 Normal distribution

2.15 Area under the curve, 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) $|Z| > 2$

2.16 Area under the curve, 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) $|Z| < 0.5$

2.17 Scores on the GRE, Part I. A college senior who took the Graduate Record Examination exam scored 620 on the Verbal Reasoning section and 670 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section was 462 with a standard deviation of 119, and the mean score for the Quantitative Reasoning was 584 with a standard deviation of 151. Suppose that both distributions are nearly normal.

(a) Write down the short-hand for these two normal distributions.

(b) What is her Z score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z scores.

(c) What do these Z scores tell you?

(d) Relative to others, which section did she do better on?

(e) Find her percentile scores for the two exams.

(f) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

(g) Explain why simply comparing her raw scores from the two sections would lead to the incorrect conclusion that she did better on the Quantitative Reasoning section.

(h) If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

2.18 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) Write down the short-hand for these two normal distributions.

(b) What are the Z scores for Leo's and Mary's finishing times? What do these Z scores tell you?

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

(d) What percent of the triathletes did Leo finish faster than in his group?

(e) What percent of the triathletes did Mary finish faster than in her group?

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

2.19 GRE scores, Part II. In Exercise 2.17 we saw two distributions for GRE scores: $N(\mu = 462, \sigma = 119)$ for the verbal part of the exam and $N(\mu = 584, \sigma = 151)$ for the quantitative part. Use this information to compute each of the following:

(a) The score of a student who scored in the 80^{th} 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.

2.20 Triathlon times, Part II. In Exercise 2.18 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.

2.21 Temperatures in LA, 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 cold are the coldest 10% of the days during June in LA?

2.22 Portfolio returns. The Capital Asset Pricing Model 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?

2.23 Temperatures in LA, Part II. Exercise 2.21 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) Write the probability model for the distribution of temperature in °C in June in LA.

(b) 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).

(c) Did you get the same answer or different answers in part (b) of this question and part (a) of Exercise 2.21? Are you surprised? Explain.

2.24 Heights of 10 year olds. 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"?

(d) 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?

2.25 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 75^{th} 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 LA.

2.26 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.[56]

(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.

2.27 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.

2.28 Find the SD. Find the standard deviation of the distribution in the following situations.

(a) MENSA is an organization whose members have IQs in the top 2% of the population. IQs are normally distributed with mean 100, and the minimum IQ score required for admission to MENSA is 132.

(b) 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.

2.29 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?

[56] 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.

2.30 SAT scores. SAT scores (out of 2400) are distributed normally with a mean of 1500 and a standard deviation of 300. Suppose a school council awards a certificate of excellence to all students who score at least 1900 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 2100? (The material covered in Section A.2 would be useful for this question.)

2.31 Scores on stats final, Part I. Below are final exam scores of 20 Introductory Statistics students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
57,	66,	69,	71,	72,	73,	74,	77,	78,	78,	79,	79,	81,	81,	82,	83,	83,	88,	89,	94

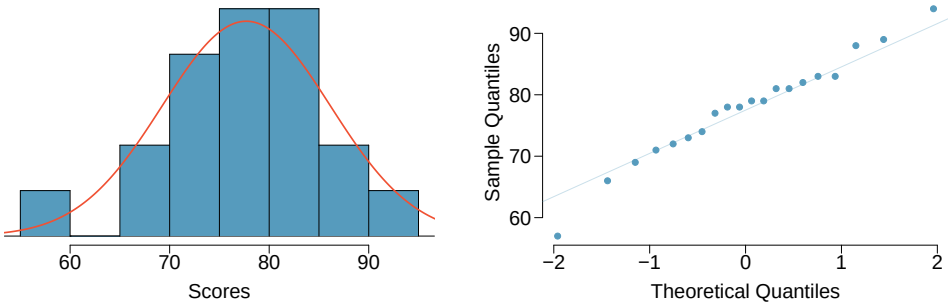
The mean score is 77.7 points. with a standard deviation of 8.44 points. Use this information to determine if the scores approximately follow the 68-95-99.7% Rule.

2.32 Heights of female college students, Part I. Below are heights of 25 female college students.

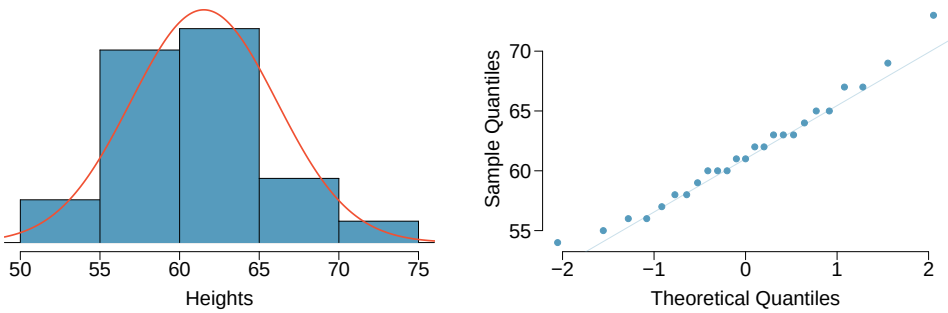
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

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.

2.33 Scores on stats final, Part II. Exercise 2.31 lists the final exam scores of 20 Introductory Statistics students. Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.



2.34 Heights of female college students, Part II. Exercise 2.32 lists the heights of 25 female college students. Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.



2.9.7 Applying the normal model

2.35 Side effects of Avandia, Part III. Exercise 2.1 introduces a study that compares the rates of serious cardiovascular problems for diabetic patients on rosiglitazone and pioglitazone treatments. The table below summarizes the results of the study.

		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) Write a set of hypotheses comparing the rates for cardiovascular problems for the two treatments.

(b) Compute the observed difference in rates for cardiovascular problems in the two treatments.

(c) This study is a suitable candidate for applying a normal distribution. If there really was no difference in the rates of cardiovascular problems and the two drugs under consideration, we can use a normal model with mean 0 and standard error 0.00084. Using this model, compute an appropriate p-value.

(d) Write a suitable conclusion based on your p-value. Use a significance level of $\alpha = 0.01$.

2.36 Crime concerns in China. A 2013 poll found that 24% of Chinese adults see crime as a very big problem, and the standard error for this estimate, which can reasonably be modeled using a normal distribution, is $SE = 1.8\%$.[57] Suppose an issue will get special attention from the Chinese government if more than 1-in-5 Chinese adults express concern on an issue.

(a) Construct hypotheses regarding whether or not crime should receive special attention by the Chinese government according to the 1-in-5 guideline.

(b) Discuss the appropriateness of using a one-sided or two-sided test for this exercise. *Consider:* for this decision process, would we care about one or both directions?

(c) Should crime receive special attention? Use a hypothesis test to justify your answer.

2.9.8 Confidence intervals

2.37 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".[58] 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.

2.38 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.[59] 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.

[57] Environmental Concerns on the Rise in China. September 19, 2013. Pew Research.
[58] The Diagnosis Difference. November 26, 2013. Pew Research.
[59] Twitter News Consumers: Young, Mobile and Educated. November 4, 2013. Pew Research.

2.39 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 Exerise 2.37 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.

2.40 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.

Chapter 3

Inference for categorical data

Chapter 3 provides a more complete framework for statistical techniques suitable for categorical data. We'll continue working with the normal model in the context of inference for proportions, and we'll also encounter a new technique and distribution suitable for working with frequency and contingency tables in Sections 3.3 and 3.4.

3.1 Inference for a single proportion

Before we get started, we'll introduce a little terminology and notation.

In the tappers-listeners study, one person tapped a tune on the table and the listener tried to guess the game. In this study, each game can be thought of as a **trial**. We could label each trial a **success** if the listener successfully guessed the tune, and we could label a trial a **failure** if the listener was unsuccessful.

Trial, success, and failure
A single event that leads to an outcome can be called a *trial*. If the trial has two possible outcomes, e.g. heads or tails when flipping a coin, we typically label one of those outcome a *success* and the other a *failure*. The choice of which outcome is labeled a success and which a failure is arbitrary, and it will not impact the results of our analyses.

When a proportion is recorded, it is common to use a 1 to represent a "success" and a 0 to represent a "failure" and then write down a **key** to communicate what each value represents. This notation is also convenient for calculations. For example, if we have 10 trials with 6 success (1's) and 4 failures (0's), the sample proportion can be computed using the mean of the zeros and ones:

$$\hat{p} = \frac{1+1+1+1+1+1+0+0+0+0}{10} = 0.6$$

Next we'll take a look at when we can apply our normal distribution framework to the distribution of the sample proportion, $\hat{p}$.

3.1.1 When the sample proportion is nearly normal

$\hat{p}$
sample proportion

p
population proportion

Conditions for when the sampling distribution of $\hat{p}$ is nearly normal
The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ 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 p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{3.1}$$

Typically we do not know the true proportion, p, so we substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value p_0 is used in place of p. Examples are presented for each of these cases in Sections 3.1.2 and 3.1.3.

TIP: Reminder on checking independence of observations
If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Or, for example, if the data come from an experiment where each user was randomly assigned to the treatment or control group and users do not interact, then the observations in each group are typically independent.

3.1.2 Confidence intervals for a proportion

According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[1] This poll included responses of 976 randomly sampled adults.

We want a confidence interval for the proportion of Americans who approve of the job the Supreme Court is doing. Our point estimate, based on a simple random sample of size $n = 976$ from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use our confidence interval formula from Section 2.8, we must first check whether the sampling distribution of $\hat{p}$ is nearly normal and calculate the standard error of the estimate.

The data are based on a simple random sample and consist of far fewer than 10% of the U.S. population, so independence is confirmed. The sample size must also be sufficiently large, which is checked via the success-failure condition: there were approximately $976 \times \hat{p} = 429$ "successes" and $976 \times (1-\hat{p}) = 547$ "failures" in the sample, both easily greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

[1] nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html

⊙ **Guided Practice 3.2** Estimate the standard error of $\hat{p} = 0.44$ using Equation (3.1). Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p. [2]

● **Example 3.3** Construct a 95% confidence interval for p, the proportion of Americans who approve of the job the Supreme Court is doing.

Using the standard error estimate from Guided Practice 3.2, the point estimate 0.44, and $z^\star = 1.96$ for a 95% confidence interval, the confidence interval can be computed as

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 0.44 \pm 1.96 \times 0.016 \quad \rightarrow \quad (0.409, 0.471)$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. At the time this poll was taken, we can say with high confidence that the job approval of the Supreme Court was below 50%.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, then the Central Limit Theorem applies, and the sampling distribution of $\hat{p}$ is well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

3.1.3 Hypothesis testing for a proportion

To apply the same normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must also be satisfied. However, in a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1 - p_0)$ are at least 10, where p_0 is the null value.

⊙ **Guided Practice 3.4** Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Ms. Toohey's opponent claimed that Ms. Toohey has *less* than 50%. Set up a hypothesis test to evaluate who is right.[3]

---

[2] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$

[3] We should run a two-sided. H_0: Ms. Toohey's support is 50%. $p = 0.50$. H_A: Ms. Toohey's support is either above or below 50%. $p \neq 0.50$.

Example 3.5 A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%. Does this provide convincing evidence for the claim of Toohey's manager at the 5% significance level?

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion, $p_0 = 0.5$: $np_0 = n(1 - p_0) = 500 \times 0.5 = 250 > 10$. With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion.

$$SE = \sqrt{\frac{p_0 \times (1 - p_0)}{n}} = \sqrt{\frac{0.5 \times (1 - 0.5)}{500}} = 0.022$$

A picture of the normal model is shown in Figure 3.1 with the p-value represented by both shaded tails. Based on the normal model, we can compute a test statistic as the Z score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.52 - 0.50}{0.022} = 0.89$$

The right tail area is 0.1867, and the p-value is $2 \times 0.1867 = 0.3734$. Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

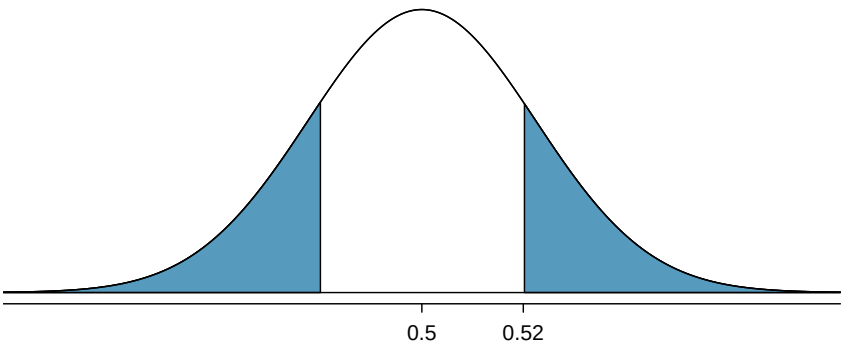


Figure 3.1: Sampling distribution of the sample proportion if the null hypothesis is true for Example 3.5. The p-value for the test is shaded.

Hypothesis test for a proportion

Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

3.1.4 Choosing a sample size when estimating a proportion

Frequently statisticians find themselves in a position to not only analyze data, but to help others determine how to most effectively collect data and also how much data should be collected. We can perform sample size calculations that are helpful in planning a study. Our task will be to identify an appropriate sample size that ensures the margin of error $ME = z^\star SE$ will be no larger than some value m. For example, we might be asked to find a sample size so the margin of error is no larger than $m = 0.04$, in which case, we write

$$z^\star SE \leq 0.04$$

Generally, we plug in a suitable value for $z^\star$ for the confidence level we plan to use, write in the formula for the standard error, and then solve for the sample size n. In the case of a single proportion, we use $\sqrt{p(1-p)/n}$ for the standard error (SE).

Example 3.6 If 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, how big of a sample is needed to ensure the margin of error is less than 0.04 using a 95% confidence level?

For a 95% confidence level, the value $z^\star$ corresponds to 1.96, and we can write the margin of error expression as follows:

$$ME = z^\star SE = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04$$

There are two unknowns in the equation: 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. The margin of error for a proportion is largest when p is 0.5, so we typically use this *worst case estimate* if no other estimate is available:

$$\begin{aligned} 1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} &\leq 0.04 \\ 1.96^2 \times \frac{0.5(1-0.5)}{n} &\leq 0.04^2 \\ 1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} &\leq n \\ 600.25 &\leq n \end{aligned}$$

We would need at least 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence. Notice that in such calculations, we always round up for the sample size!

As noted in the example, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

⊙ **Guided Practice 3.7** A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.[4]

(a) There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.

(b) The sample sizes vary widely. Which of the three would you suggest using? What would influence your choice?

⊙ **Guided Practice 3.8** A recent estimate of Congress' approval rating was 17%.[5] If we were to conduct a new poll and wanted an estimate with a margin of error smaller than about 0.04 with 95% confidence, how big of a sample should we use?[6]

3.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions $(p_1 - p_2)$ using the normal model. In this section we consider three such examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, a company weighs whether they should switch to a higher quality parts manufacturer. In the last example, we examine the cancer risk to dogs from the use of yard herbicides.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions; as before, these conditions relate to independence of observations and checking for sufficiently large sample size. Finally, we compute the estimate's standard error and apply our inferential framework.

[4] (a) For the 1.7% estimate of p, we estimate the appropriate sample size as follows:

$$1.65 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \leq 0.02 \qquad \rightarrow \qquad n \geq 113.7$$

Using the estimate from the first model, we would suggest examining 114 tires (round up!). A similar computation can be accomplished using 0.062 and 0.013 for p: 396 and 88.

(b) We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. (Answers will vary.)

[5] www.gallup.com/poll/155144/Congress-Approval-June.aspx

[6] We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \leq 0.04 \qquad \rightarrow \qquad n \geq 338.8$$

A sample size of 339 or more would be reasonable.

3.2.1 Sample distribution of the difference of two proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. 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.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal
The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when

- each proportion separately follows a normal model, and
- the two samples are independent of each other.

The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{3.9}$$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

3.2.2 Intervals and tests for $p_1 - p_2$

In the setting of confidence intervals, the sample proportions are used to verify the success-failure condition and also compute standard error, just as was the case with a single proportion.

Example 3.10 The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[7]

> 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 order given above, or the two statements were reversed. Table 3.2 shows the results of this experiment. Create and interpret a 90% confidence interval of the difference in approval.

First the conditions must be verified. Because each group is a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample. Because all conditions are met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13$$

[7] 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.

	Sample size (n_i)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Table 3.2: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

The standard error may be computed from Equation (3.9) using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025$$

For a 90% confidence interval, we use $z^\star = 1.65$:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 0.13 \pm 1.65 \times 0.025 \quad \rightarrow \quad (0.09, 0.17)$$

We are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still fluid on the health insurance mandate.

Guided Practice 3.11 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. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[8]

Example 3.12 The quality control engineer from Guided Practice 3.11 collects a sample of gears, examining 1000 gears from each company and finds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Guided Practice 3.11 using a significance level of 5%.

First, we check the conditions. The sample is 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. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, $0.958 - 0.899 = 0.059$, can be said to come from a nearly normal distribution.

[8] H_0: The higher quality gears will pass inspection no more than 3% more frequently than the standard quality gears. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher quality gears will pass inspection more than 3% more often than the standard quality gears. $p_{highQ} - p_{standard} > 0.03$.

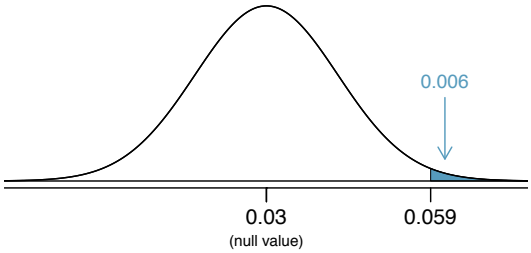


Figure 3.3: Distribution of the test statistic if the null hypothesis was true. The p-value is represented by the shaded area.

The standard error can be found using Equation (3.9):

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114$$

In this hypothesis test, the sample proportions were used. We will discuss this choice more in Section 3.2.3.

Next, we compute the test statistic and use it to find the p-value, which is depicted in Figure 3.3.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher quality gears actually do pass inspection more than 3% as often as the currently used gears. Based on these results, management will approve the switch to the new supplier.

3.2.3 Hypothesis testing when $H_0 : p_1 = p_2$

Here we use a new example to examine a special estimate of standard error when $H_0 : p_1 = p_2$. We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.[9] Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table 3.4.

[9]Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. Case-Control Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.

	cancer	no cancer
2,4-D	191	304
no 2,4-D	300	641

Table 3.4: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

Guided Practice 3.13 Is this study an experiment or an observational study?[10]

Guided Practice 3.14 Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups.[11]

Example 3.15 Are the conditions met to use the normal model and make inference on the results?

(1) It is unclear whether this is a random sample. However, if we believe the dogs in both the cancer and no cancer groups are representative of each respective population and that the dogs in the study do not interact in any way, then we may find it reasonable to assume independence between observations. (2) The success-failure condition holds for each sample.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

In your hypotheses for Guided Practice 3.14, the null is that the proportion of dogs with exposure to 2,4-D is the same in each group. The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. To identify the p-value for this test, we first check conditions (Example 3.15) and compute the standard error of the difference:

$$SE = \sqrt{\frac{p_c(1-p_c)}{n_c} + \frac{p_n(1-p_n)}{n_n}}$$

In a hypothesis test, the distribution of the test statistic is always examined as though the null hypothesis is true, i.e. in this case, $p_c = p_n$. The standard error formula should reflect this equality in the null hypothesis. We will use p to represent the common rate of dogs that are exposed to 2,4-D in the two groups:

$$SE = \sqrt{\frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_n}}$$

[10] The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the *control group*, which is a term usually seen in experiments.

[11] Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reflect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.
H_0: the proportion of dogs with exposure to 2,4-D is the same in "cancer" and "no cancer" dogs, $p_c - p_n = 0$.
H_A: dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

We don't know the exposure rate, p, but we can obtain a good estimate of it by *pooling* the results of both samples:

$$\hat{p} = \frac{\text{\# of "successes"}}{\text{\# of cases}} = \frac{191 + 304}{191 + 300 + 304 + 641} = 0.345$$

This is called the **pooled estimate** of the sample proportion, and we use it to compute the standard error when the null hypothesis is that $p_1 = p_2$ (e.g. $p_c = p_n$ or $p_c - p_n = 0$). We also typically use it to verify the success-failure condition.

Pooled estimate of a proportion

When the null hypothesis is $p_1 = p_2$, it is useful to find the pooled estimate of the shared proportion:

$$\hat{p} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

TIP: Use the pooled proportion estimate when $H_0 : p_1 = p_2$

When the null hypothesis suggests the proportions are equal, we use the pooled proportion estimate ($\hat{p}$) to verify the success-failure condition and also to estimate the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \tag{3.16}$$

⊙ **Guided Practice 3.17** Using Equation (3.16), $\hat{p} = 0.345$, $n_1 = 491$, and $n_2 = 945$, verify the estimate for the standard error is $SE = 0.026$. Next, complete the hypothesis test using a significance level of 0.05. Be certain to draw a picture, compute the p-value, and state your conclusion in both statistical language and plain language.[12]

[12] Compute the test statistic:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up $Z = 2.58$ in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: $1 - 0.9951 = 0.0049$. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

3.3 Testing for goodness of fit using chi-square (special topic)

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution. (Background on the geometric distribution is not necessary.)

Each of these scenarios can be addressed using the same statistical test: a chi-square test.

In the first case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Table 3.5, 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

Table 3.5: 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.

A second application, assessing the fit of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for the years 1990-2011 are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

3.3.1 Creating a test statistic for one-way tables

Example 3.18 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 3.19 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 Table 3.6.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Table 3.6: 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.

3.3.2 The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

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.

Our strategy will be to first compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50$$

[13] Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown – though not here – that the square root of the count is a better way to standardize the count differences.

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

Black
$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

Hispanic
$$Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39$$

Other
$$Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they – as a group – tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58$$

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual – e.g. a standardized difference of 2.5 – will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

X^2
chi-square test statistic

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \cdots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4}$$

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 3.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a *chi-square distribution*. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

3.3.3 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 the 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 3.20** Figure 3.7 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?[14]

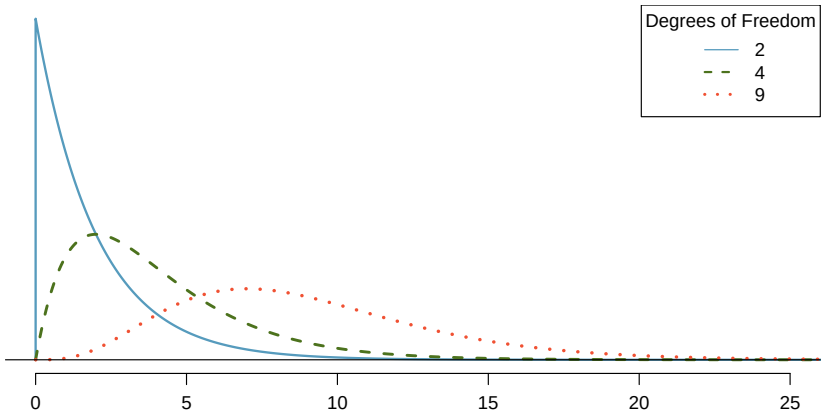


Figure 3.7: Three chi-square distributions with varying degrees of freedom.

Figure 3.7 and Guided Practice 3.20 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.

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 Table 3.8. A more complete table is presented in Appendix C.3 on page 344. Using this table, we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important quality of this table: 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	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

Table 3.8: A section of the chi-square table. A complete table is in Appendix C.3 on page 344.

[14](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 skewed for $df = 2$, and then the distributions become more symmetric for the larger degrees of freedom $df = 4$ and $df = 9$. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Example 3.21 Figure 3.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 3.8 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 3.9(a) has area 0.1.

Example 3.22 We rarely observe the *exact* value in the table. For instance, Figure 3.9(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3, which does not fall in Table 3.8. 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 3.9(b) is between 0.1 and 0.2.

Example 3.23 Figure 3.9(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.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than 0.3*.

Guided Practice 3.24 Figure 3.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

Guided Practice 3.25 Figure 3.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Guided Practice 3.26 Figure 3.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

3.3.4 Finding a p-value for a chi-square distribution

In Section 3.3.2, we identified a new test statistic (X^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 X^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 ($X^2 = 5.89$) 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 X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of bins.

[15] The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.
[16] The area is between 0.02 and 0.05.
[17] Between 0.02 and 0.05.

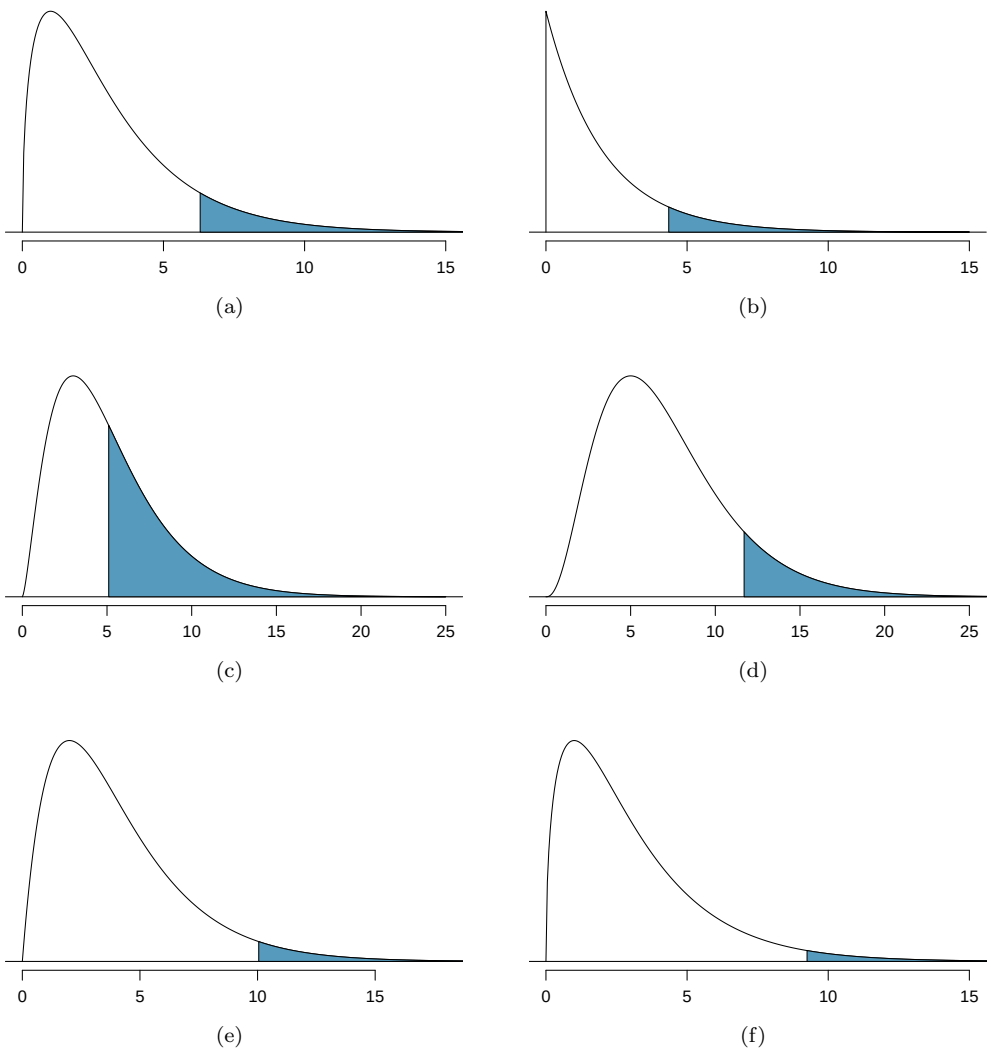


Figure 3.9: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

Example 3.27 How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

In the jurors example, there were $k = 4$ categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^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 apply the chi-square distribution for X^2. 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 apply the chi-square model to the test statistic, $X^2 = 5.89$.

Example 3.28 If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

The chi-square distribution and p-value are shown in Figure 3.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix C.3 or the short table on page 137, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

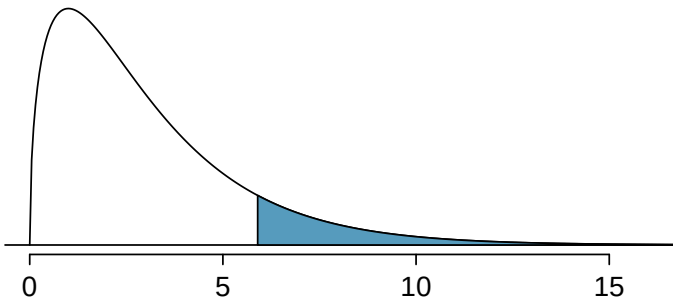


Figure 3.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

Chi-square 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. Call 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:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

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 X^2 would provide greater evidence against the null hypothesis.

TIP: Conditions for the chi-square test

There are three conditions that must be checked before performing a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Degrees of freedom We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one-proportion methods introduced in Section 3.1.

3.3.5 Evaluating goodness of fit for a distribution

We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model fits a data set. Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as `Up` or `Down` (`D`) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each `Up` day:

Change in price	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
Outcome	Up	D	Up	D	Up	Up	D	D	D	Up
Days to Up	1	-	2	-	2	1	-	-	-	4

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the k^{th} trial to observe the first success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before

we observed our next `Up` trading day, and two more for the third `Up` day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Table 3.11 shows the number of waiting days for a positive trading day during 1990-2011 for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

Table 3.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock exchange. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a *geometric distribution*. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an `Up` day is observed. Under this hypothesis, the number of days until an `Up` day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an `Up` day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 from 1990 to 2011 and summarize the waiting times in Table 3.12 and Figure 3.13. The S&P500 was positive on 53.2% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have *binned* together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the *Observed* row in Table 3.12, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Table 3.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each bin, then (2) multiplying each null proportion by the total count to obtain the expected

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

Table 3.12: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model ($P(D) = (1 - 0.532)^{D-1}(0.532)$) and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 = 11.65\%$ of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

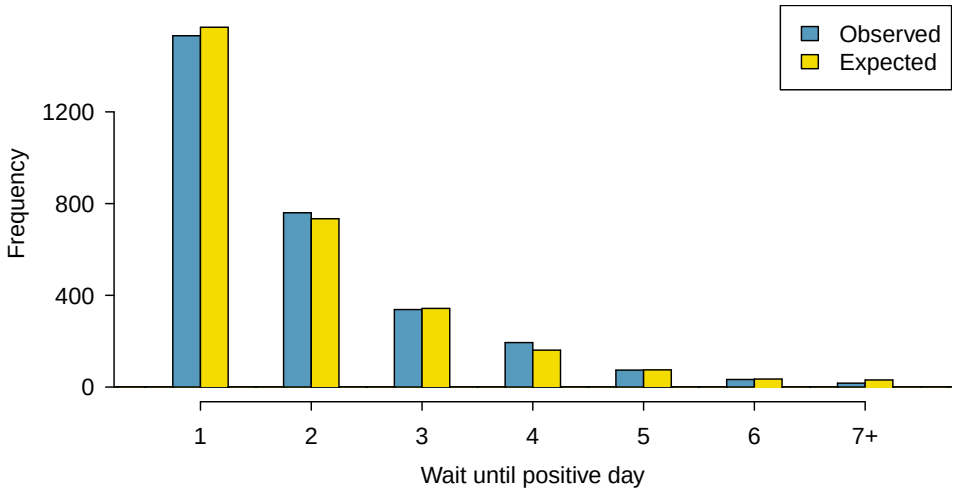


Figure 3.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Example 3.29 Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Table 3.12.

Guided Practice 3.30 Table 3.12 provides a set of count data for waiting times ($O_1 = 1532$, $O_2 = 760$, ...) and expected counts under the geometric distribution ($E_1 = 1569$, $E_2 = 734$, ...). Compute the chi-square test statistic, X^2.[18]

Guided Practice 3.31 Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

Example 3.32 If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

Figure 3.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix C.3, we find that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

[18] $X^2 = \frac{(1532-1569)^2}{1569} + \frac{(760-734)^2}{734} + \cdots + \frac{(17-31)^2}{31} = 15.08$

[19] There are $k = 7$ groups, so we use $df = k - 1 = 6$.

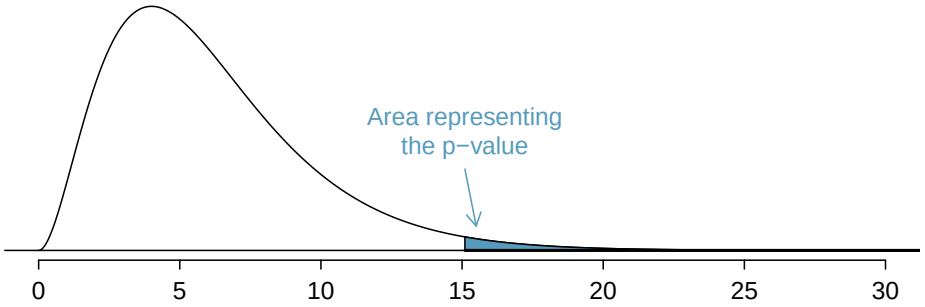


Figure 3.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

Example 3.33 In Example 3.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

3.4 Testing for independence in two-way tables (special topic)

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. Table 3.15 shows an example of 10,000 queries split into three algorithm groups.[20] 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

Table 3.15: Google experiment breakdown of test subjects into three search groups.

Example 3.34 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.

[20] 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.

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) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. 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.

Table 3.16 provides the results from the experiment. These data are very similar to the count data in Section 3.3. 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

Table 3.16: 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? That is, are they dependent (versus independent)?

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 `search` variable is independent of the `algorithm` variable.

3.4.1 Expected counts in two-way tables

Example 3.35 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 3.36** Using the same rationale described in Example 3.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 3.35 and Guided Practice 3.36. These expected counts were used to construct Table 3.17, which is the same as Table 3.16, 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

Table 3.17: The observed counts and the **(expected counts)**.

The examples and guided practice 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.

Computing expected counts in a two-way table

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i,\ \text{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

[21] We would expect $0.7078 * 2500 = 1769.5$. It is okay that this is a fraction.

3.4.2 The chi-square test 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	$\dfrac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\dfrac{(3511 - 3539)^2}{3539} = 0.222$
Row 1, Col 2	$\dfrac{(1749 - 1769.5)^2}{1769.5} = 0.237$
$\vdots$	$\vdots$
Row 2, Col 3	$\dfrac{(682 - 730.5)^2}{730.5} = 3.220$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1})$$

In our example, the degrees of freedom parameter 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 $X^2 = 6.12$ 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 3.18.

Computing degrees of freedom for a two-way table

When applying 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.

TIP: 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 3.2.

[22] Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.

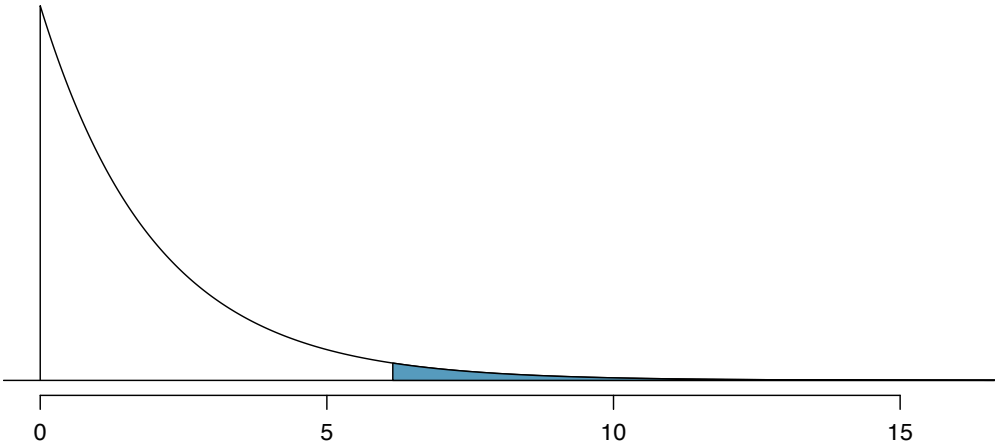


Figure 3.18: Computing the p-value for the Google hypothesis test.

		Congress		
	Obama	Democrats	Republicans	Total
Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

Table 3.19: Pew Research poll results of a March 2012 poll.

Example 3.37 Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Looking in Appendix C.3 on page 344, we examine the row corresponding to 2 degrees of freedom. The test statistic, $X^2 = 6.120$, falls between the fourth and fifth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of $\alpha = 0.05$ and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Example 3.38 Table 3.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

H_0: There is no difference in approval ratings between the three groups.

H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23] See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 3.19 are approximate.

⊙ **Guided Practice 3.39** A chi-square test for a two-way table may be used to test the hypotheses in Example 3.38. As a first step, compute the expected values for each of the six table cells.[24]

⊙ **Guided Practice 3.40** Compute the chi-square test statistic.[25]

⊙ **Guided Practice 3.41** Because there are 2 rows and 3 columns, the degrees of freedom for the test is $df = (2-1) \times (3-1) = 2$. Use $X^2 = 106.4$, $df = 2$, and the chi-square table on page 344 to evaluate whether to reject the null hypothesis.[26]

[24]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): $\frac{2119 \times 1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104 \times 1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

[25]For each cell, compute $\frac{(\text{obs}-\text{exp})^2}{exp}$. For instance, the first row and first column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \cdots + 34.0 = 106.4$.

[26]The test statistic is larger than the right-most column of the $df = 2$ row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans' approval has differences among Democrats in Congress, Republicans in Congress, and the president.

3.5 Exercises

3.5.1 Inference for a single proportion

3.1 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.

3.2 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.[27]

(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.

3.3 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.

3.4 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.[28]

(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 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.

[27] A. Vaughn. "Poll finds young adults optimistic, but not about money". In: *Los Angeles Times* (2011).
[28] Demos.org. "The State of Young America: The Poll". In: (2011).

3.5 Prop 19 in California. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[29]

(a) We are 95% confident that between 62% and 78% of the California voters in this sample support Prop 19.

(b) We are 95% confident that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.

(c) If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.

(d) In order to decrease the margin of error to 4%, 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 California voters between the ages of 18 and 34 support Prop 19.

3.6 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.[30]

(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%.

3.7 Fireworks on July 4^{th}. In late June 2012, Survey USA published results of a survey stating that 56% of the 600 randomly sampled Kansas residents planned to set off fireworks on July 4^{th}. Determine the margin of error for the 56% point estimate using a 95% confidence level.[31]

3.8 Elderly drivers. In January 2011, 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.[32]

(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?

[29] Survey USA, Election Poll #16804, data collected July 8-11, 2010.
[30] Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.
[31] Survey USA, News Poll #19333, data collected on June 27, 2012.
[32] Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.

3.9 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.

3.10 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".[33]

(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.

3.11 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[34]

(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?

[33] Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.

[34] studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

3.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[35]

(a) Is 48% 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?

3.13 Public option, Part I. A *Washington Post* article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More specifically, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." (6% responded with "other".) There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.[36]

(a) A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

(b) Would you expect a confidence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

3.14 The Civil War. A national survey conducted in 2011 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.[37]

(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.

3.15 Browsing on the mobile device. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.[38]

(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.[39] 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.

---

[35] National Opinion Research Center, General Social Survey, 2010.

[36] D. Balz and J. Cohen. "Most support public option for health insurance, poll finds". In: *The Washington Post* (2009).

[37] Pew Research Center Publications, Civil War at 150: Still Relevant, Still Divisive, data collected between March 30 - April 3, 2011.

[38] Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.

[39] S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: *M.I.C Gadget* (2012).

3.16 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.[40]

(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.

3.17 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 able to detect the difference between diet and regular soda, in other words, are the results significantly better than just random guessing?

(b) Interpret the p-value in this context.

3.18 Is college worth it? Part II. Exercise 3.16 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?

3.19 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?

3.20 Legalize Marijuana, Part II. As discussed in Exercise 3.12, the 2010 General Social Survey reported a sample where about 48% 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 ?

3.21 Public option, Part II. Exercise 3.13 presents the results of a poll evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

[40] Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.

3.22 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?

3.5.2 Difference of two proportions

3.23 Social experiment. 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		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	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.

3.24 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.[41]

	control	treatment
alive	4	24
dead	30	45

A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate 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?

[41] B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

3.25 Gender and color preference. A 2001 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.[42]

(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.

3.26 The Daily Show. A 2010 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.[43]

(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).

3.27 Public Option, Part III. Exercise 3.13 presents the results of a poll evaluating support for the health care public option plan in 2009. 70% of 819 Democrats and 42% of 783 Independents support the public option.

(a) Calculate a 95% confidence interval for the difference between ($p_D - p_I$) 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 public option than the Independent.

[42] L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: *Personality and Individual Differences* 31.8 (2001), pp. 1375–1379.

[43] The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.

3.28 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.[44]

3.29 Offshore drilling, Part I. A 2010 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.[45]

(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

3.30 Sleep deprivation, CA vs. OR, Part II. Exercise 3.28 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?

3.31 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 3.29.

	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.

[44] CDC, Perceived Insufficient Rest or Sleep Among Adults — United States, 2008.
[45] Survey USA, Election Poll #16804, data collected July 8-11, 2010.

3.32 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, interviewed by telephone November 7-10, 2010, 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.[46]

		Party Affiliation		
		Republican	Democrat	Independent
Answer	Should	264	299	351
	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 I or a Type II error? Explain.

3.33 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.[47]

		Transportation Professionals			
	Control	Pilots	Truck Drivers	Train Operators	Bux/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.

[46] S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: *CBS News* (2010).
[47] National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers Sleep, 2012.

3.34 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).[48]

		Autism		
		Autism	Typical development	Total
Periconceptional prenatal vitamin	No vitamin	111	70	181
	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.[49]

3.35 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.[50]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

3.36 Diabetes and unemployment. A 2012 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.[51]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of incidence of diabetes and employment status.

(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.

---

[48] R.J. Schmidt et al. "Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism". In: *Epidemiology* 22.4 (2011), p. 476.

[49] R.C. Rabin. "Patterns: Prenatal Vitamins May Ward Off Autism". In: *New York Times* (2011).

[50] S. Lockman et al. "Response to antiretroviral therapy after a single, peripartum dose of nevirapine". In: *Obstetrical & gynecological survey* 62.6 (2007), p. 361.

[51] Gallup Wellbeing, Employed Americans in Better Health Than the Unemployed, data collected Jan. 2, 2011 - May 21, 2012.

3.5.3 Testing for goodness of fit using chi-square

3.37 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.

3.38 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 $X^2 = 10$ 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.

3.39 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.

3.40 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[52]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1)	38%	37%
Humans evolved, but God had no part in process (2)	16%	12%
God created humans in present form (3)	40%	45%
Other / No opinion (4)	6%	6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)

[52] Four in 10 Americans Believe in Strict Creationism, December 17, 2010, http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx.

3.5.4 Testing for independence in two-way tables

3.41 Offshore drilling, Part III. The table below summarizes a data set we first encountered in Exercise 3.29 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

3.42 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.[53]

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical depression	Yes	670	373	905	564	95	2,607
	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 $X^2 = 20.93$. 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.[54] Do you agree with this statement? Explain your reasoning.

[53] M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: *Archives of internal medicine* 171.17 (2011), p. 1571.

[54] A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: *New York Times* (2011).

3.43 Privacy on Facebook. 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.[55]

		Gender		
		Male	Female	Total
Response	Yes	288	378	666
	No	61	62	123
	Not sure	10	7	17
	Total	359	447	806

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

3.44 Shipping holiday gifts. A December 2010 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
Shipping Method	USPS	72	(81)	97	(102)	76	(62)	245
	UPS	52	(53)	76	(68)	34	(41)	162
	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?

[55] Survey USA, News Poll #17960, data collected February 16-17, 2011.

Chapter 4

Inference for numerical data

Chapters 2 and 3 introduced us to inference for proportions using the normal model, and in Section 3.3, we encountered the chi-square distribution, which is useful for working with categorical data with many levels. In this chapter, our focus will be on numerical data, where we will encounter two more distributions: the t distribution (looks a lot like the normal distribution) and the F distribution. Our general approach will be:

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 hypothesis and confidence interval techniques from Chapter 2 using the distribution from step 2.

4.1 One-sample means with the t distribution

The sampling distribution associated with a sample mean or difference of two sample means is, if certain conditions are satisfied, nearly normal. However, this becomes more complex when the sample size is small, where *small* here typically means a sample size smaller than 30 observations. For this reason, we'll use a new distribution called the t distribution that will often work for both small and large samples of numerical data.

4.1.1 Two examples using the normal distribution

Before we get started with the t distribution, let's take a look at two applications where it is okay to use the normal model for the sample mean. For the case of a single mean, the standard error of the sample mean can be calculated as

$$SE = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size. Generally we use the sample standard deviation, denoted by s, in place of the population standard deviation when we compute the standard error:

$$SE \approx \frac{s}{\sqrt{n}}$$

If we look at this formula, there are some characteristics that we can think about intuitively.

- If we examine the standard error formula, we would see that a larger s corresponds to a larger SE. This makes intuitive sense: if the data are more volatile, then we'll be less certain of the location of the true mean, so the standard error should be bigger. On the other hand, if the observations all fall very close together, then s will be small, and the sample mean should be a more precise estimate of the true mean.
- In the formula, the larger the sample size n, the smaller the standard error. This matches our intuition: we expect estimates to be more precise when we have more data, so the standard error SE should get smaller when n gets bigger.

As we did with proportions, we'll also need to check a few conditions before using the normal model. We'll forgo describing those details until later this section, but these conditions have been verified for the two examples below.

Example 4.1 We've taken a random sample of 100 runners from a race called the Cherry Blossom Run in Washington, DC, which was a race with 16,924 participants.[1] The sample data for the 100 runners is summarized in Table 4.1, histograms of the run time and age of participants are in Figure 4.2, and summary statistics are available in Table 4.3. Create a 95% confidence interval for the average time it takes runners in the Cherry Blossom Run to complete the race.

We can use the same confidence interval formula for the mean that we used for a proportion:

$$\text{point estimate} \ \pm \ 1.96 \times SE$$

In this case, the best estimate of the overall mean is the sample mean, $\bar{x} = 95.61$ minutes. The standard error can be calculated using sample standard deviation ($s = 15.78$), the sample size ($n = 100$), and the standard error formula:

$$SE = \frac{s}{\sqrt{n}} = \frac{15.78}{\sqrt{100}} = 1.578$$

Finally, we can calculate a 95% confidence interval:

$$\text{point estimate} \ \pm \ z^\star \times SE \quad \rightarrow \quad 95.61 \pm 1.96 \times 1.578 \quad \rightarrow \quad (92.52, 98.70)$$

We are 95% confident that the average time for all runners in the 2012 Cherry Blossom Run is between 92.52 and 98.70 minutes.

ID	time	age	gender	state
1	88.31	59	M	MD
2	100.67	32	M	VA
3	109.52	33	F	VA
⋮	⋮	⋮	⋮	⋮
100	89.49	26	M	DC

Table 4.1: Four observations for the `run10Samp` data set, which represents a simple random sample of 100 runners from the 2012 Cherry Blossom Run.

[1] See www.cherryblossom.org.

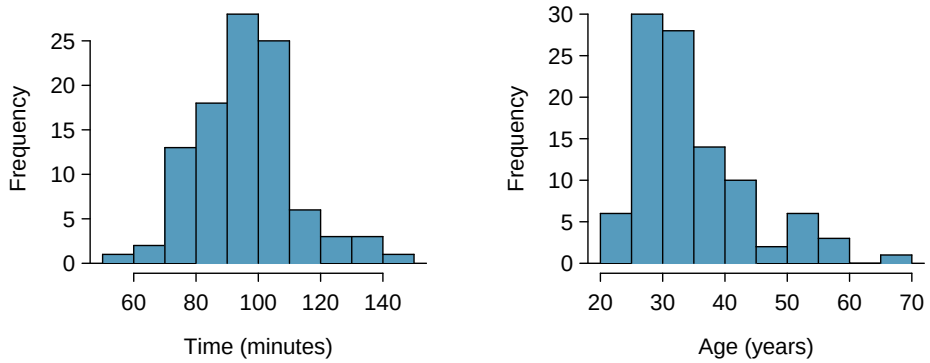


Figure 4.2: Histograms of `time` and `age` for the sample Cherry Blossom Run data. The average time is in the mid-90s, and the average age is in the mid-30s. The age distribution is moderately skewed to the right.

	`time`	`age`
sample mean	95.61	35.05
sample median	95.37	32.50
sample st. dev.	15.78	8.97

Table 4.3: Point estimates and parameter values for the `time` variable.

Guided Practice 4.2 Use the data to calculate a 90% confidence interval for the average age of participants in the 2012 Cherry Blossom Run. The conditions for applying the normal model have already been verified.[2]

Example 4.3 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 sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? The conditions necessary for applying the normal model have been checked and are satisfied.

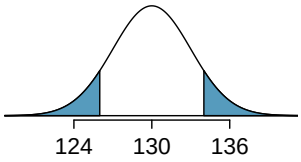
The question has been framed in terms of two possibilities: the nutrition label accurately lists the correct average calories per bag of chips or it does not, which may be framed as a hypothesis test:

H_0: The average is listed correctly. $\mu = 130$

H_A: The nutrition label is incorrect. $\mu \neq 130$

The observed average is $\bar{x} = 134$ and the standard error may be calculated as $SE = \frac{17}{\sqrt{35}} = 2.87$. First, we draw a picture summarizing this scenario.

[2] As before, we identify the point estimate, $\bar{x} = 35.05$, and the standard error, $SE = 8.97/\sqrt{100} = 0.897$. Next, we apply the formula for a 90% confidence interval, which uses $z^\star = 1.65$: $35.05 \pm 1.65 \times 0.897 \rightarrow (33.57, 36.53)$. We are 90% confident that the average age of all participants in the 2012 Cherry Blossom Run is between 33.57 and 36.53 years.



We can compute a test statistic as the Z score:

$$Z = \frac{134 - 130}{2.87} = 1.39$$

The upper-tail area is 0.0823, so the p-value is $2 \times 0.0823 = 0.1646$. Since the p-value is larger than 0.05, we do not reject the null hypothesis. That is, there is not enough evidence to show the nutrition label has incorrect information.

The normal model works well when the sample size is larger than about 30. For smaller sample sizes, we run into a problem: our estimate of s, which is used to compute the standard error, isn't as reliable when the sample size is small. To solve this problem, we'll use a new distribution: the t distribution.

4.1.2 Introducing the t distribution

A t distribution, shown as a solid line in Figure 4.4, has a bell shape that looks very similar to a normal distribution (dotted line). However, its tails are thicker, which means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[3] When our sample is small, the value s used to compute the standard error isn't very reliable. The extra thick tails of the t distribution are exactly the correction we need to resolve this problem.

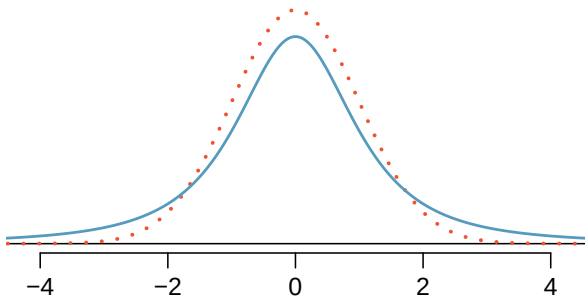


Figure 4.4: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

The t distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell-shaped t distribution. Several t distributions are shown in Figure 4.5 with various degrees of freedom. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution.

[3]The standard deviation of the t distribution is actually a little more than 1. However, it is useful to always think of the t distribution as having a standard deviation of 1 in all of our applications.

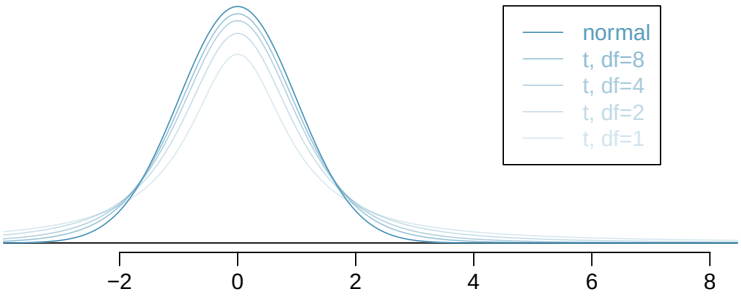


Figure 4.5: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

Degrees of freedom (df)
The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution, e.g. see Figure 4.5. In Section 4.1.3, we relate degrees of freedom to sample size.

We will find it very useful to become familiar with the t distribution, because it plays a very similar role to the normal distribution during inference for numerical data. We use a **t table**, partially shown in Table 4.6, in place of the normal probability table for small sample numerical data. A larger table is presented in Appendix C.2 on page 342. Alternatively, we could use statistical software to get this same information.

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
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
	⋮	⋮	⋮	⋮	⋮	
	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
	⋮	⋮	⋮	⋮	⋮	
	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

Table 4.6: 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**.

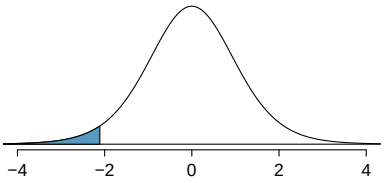


Figure 4.7: The t distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

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 Table 4.6. 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 4.4 What proportion of the t distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 4.7 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. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

Example 4.5 A t distribution with 20 degrees of freedom is shown in the left panel of Figure 4.8. 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.

Example 4.6 A t distribution with 2 degrees of freedom is shown in the right panel of Figure 4.8. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: $df = 2$. Next, find the columns that capture 3; because $2.92 < 3 < 4.30$, we use the second and third columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

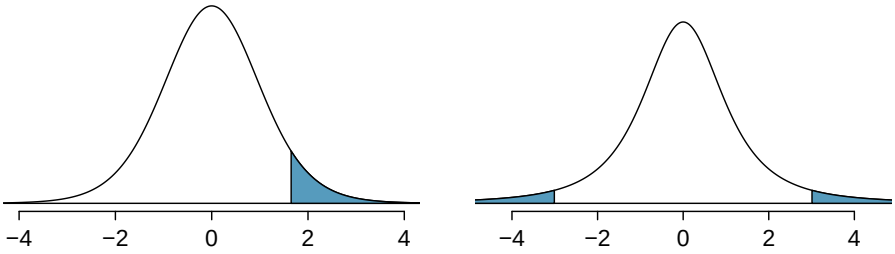


Figure 4.8: Left: The t distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

⊙ **Guided Practice 4.7** What proportion of the t distribution with 19 degrees of freedom falls above -1.79 units?[4]

4.1.3 Applying the t distribution to the single-mean situation

When estimating the mean and standard error from a sample of numerical data, the t distribution is a little more accurate than the normal model. This is true for both small and large samples, though the benefits for larger samples are limited.

Using the t distribution
Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

Before applying the t distribution for inference about a single mean, we check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if the data are from an experiment or random process, we carefully check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, usually in the form of prominent outliers, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal. However, if the sample size is somewhat large, then we can relax this condition, e.g. moderate skew is acceptable when the sample size is 30 or more, and strong skew is acceptable when the size is about 60 or more.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t distribution with $n - 1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t distribution

[4] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.

with $df = 19 - 1 = 18$ degrees of freedom and proceed in the same way as we did in Chapter 3, except that *now we use the t table.*

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.

4.1.4 One sample t confidence intervals

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 occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

Figure 4.9: A Risso's dolphin.
Photo by Mike Baird (http://www.bairdphotos.com/).

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[5] The data are summarized in Table 4.10. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 4.10: 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).

[5]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. Data reference: 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.

Example 4.8 Are the independence and normality conditions satisfied for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. Ideally we would see a visualization of the data to check for skew and outliers. However, we can instead examine the summary statistics in Table 4.10, which do not suggest any skew or outliers. All observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

In the normal model, we used $z^\star$ and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t distribution:

$$\bar{x} \pm t^\star_{df} \times SE$$

$t^\star_{df}$
Multiplication factor for t conf. interval

The sample mean and estimated standard error are computed just as in our earlier examples that used the normal model ($\bar{x} = 4.4$ and $SE = s/\sqrt{n} = 0.528$). The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t distribution with df degrees of freedom. Before determining this cutoff, we will first need the degrees of freedom.

In our current example, we should use the t distribution with $df = n - 1 = 19 - 1 = 18$ degrees of freedom. Then identifying $t^\star_{18}$ is similar to how we found $z^\star$:

- For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t distribution is between -$t^\star_{18}$ and $t^\star_{18}$.
- We look in the t table on page 167, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

Generally the value of $t^\star_{df}$ is slightly larger than what we would get under the normal model with $z^\star$.

Finally, we can substitute all the values into the confidence interval equation to create the 95% confidence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \pm t^\star_{18} \times SE \quad \rightarrow \quad 4.4 \pm 2.10 \times 0.528 \quad \rightarrow \quad (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram, which is considered extremely high.

Finding a t confidence interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \pm t^\star_{df} \times SE$$

where $\bar{x}$ is the sample mean, $t^\star_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample. The normality condition may be relaxed for larger sample sizes.

⊙ **Guided Practice 4.9** The FDA's webpage provides some data on mercury content of fish.[6] 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. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[7]

● **Example 4.10** Estimate the standard error of the sample mean using the data summaries in Guided Practice 4.9. If we are to use the t distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^\star_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: $df = n - 1 = 14$.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row $df = 14$, we identify $t^\star_{14} = 1.76$.

⊙ **Guided Practice 4.11** Using the results of Guided Practice 4.9 and Example 4.10, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[8]

4.1.5 One sample t tests

Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Run, comparing runners in 2006 and 2012. Technological advances in shoes, training, and diet might suggest runners would be faster in 2012. 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.

The average time for all runners who finished the Cherry Blossom Run in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine using data from 100 participants in the 2012 Cherry Blossom Run whether runners in this race are getting faster or slower, versus the other possibility that there has been no change.

⊙ **Guided Practice 4.12** What are appropriate hypotheses for this context?[9]

⊙ **Guided Practice 4.13** The data come from a simple random sample from less than 10% of all participants, so the observations are independent. However, should we be worried about skew in the data? A histogram of the differences was shown in the left panel of Figure 4.2 on page 165. [10]

With independence satisfied and skew not a concern, we can proceed with performing a hypothesis test using the t distribution.

[6] http://www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

[7] There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red flags for the normal model based on this (limited) information, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

[8] $\bar{x} \pm t^\star_{14} \times SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

[9] H_0: The average 10 mile run time was the same for 2006 and 2012. $\mu = 93.29$ minutes. H_A: The average 10 mile run time for 2012 was *different* than that of 2006. $\mu \neq 93.29$ minutes.

[10] With a sample of 100, we should only be concerned if there is extreme skew. The histogram of the data suggest, at worst, slight skew.

⊙ **Guided Practice 4.14** The sample mean and sample standard deviation are 95.61 and 15.78 minutes, respectively. Recall that the sample size is 100. What is the p-value for the test, and what is your conclusion?[11]

When using a t distribution, we use a T score (same as Z score)
To help us remember to use the t distribution, we use a T to represent the test statistic, and we often call this a **T score**. The Z score and T score are computed in the exact same way and are conceptually identical: each represents how many standard errors the observed value is from the null value.

4.2 Paired data

Are textbooks actually cheaper online? Here we compare the price of textbooks at the University of California, Los Angeles' (UCLA's) bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses.[12] A portion of the data set is shown in Table 4.11.

	dept	course	ucla	amazon	diff
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

Table 4.11: Six cases of the `textbooks` data set.

4.2.1 Paired observations

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**.

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.

[11] With the conditions satisfied for the t distribution, we can compute the standard error ($SE = 15.78/\sqrt{100} = 1.58$ and the T *score*: $T = \frac{95.61-93.29}{1.58} = 1.47$. (There is more on this after the guided practice, but a T score and Z score are basically the same thing.) For $df = 100 - 1 = 99$, we would find $T = 1.47$ to fall between the first and second column, which means the p-value is between 0.05 and 0.10 (use $df = 90$ and consider two tails since the test is two-sided). Because the p-value is greater than 0.05, we do not reject the null hypothesis. That is, the data do not provide strong evidence that the average run time for the Cherry Blossom Run in 2012 is any different than the 2006 average.

[12] When a class had multiple books, only the most expensive text was considered.

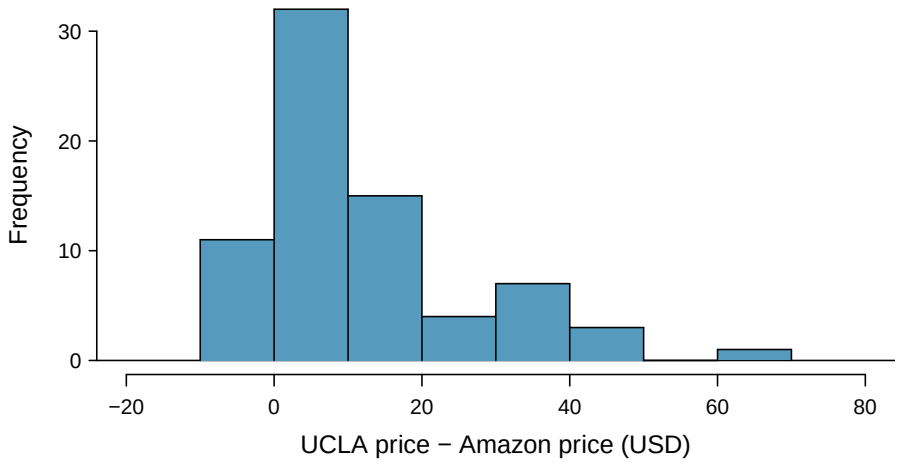


Figure 4.12: Histogram of the difference in price for each book sampled. These data are strongly skewed.

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 difference in prices, which is represented as the `diff` variable in the `textbooks` data. Here the differences are taken as

$$\text{UCLA price} - \text{Amazon price}$$

for each book. 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 4.12. Using differences between paired observations is a common and useful way to analyze paired data.

⊙ **Guided Practice 4.15** The first difference shown in Table 4.11 is computed as $27.67 - 27.95 = -0.28$. Verify the differences are calculated correctly for observations 2 and 3.[13]

4.2.2 Inference 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}
73	12.76	14.26

Table 4.13: Summary statistics for the price differences. There were 73 books, so there are 73 differences.

[13] Observation 2: $40.59 - 31.14 = 9.45$. Observation 3: $31.68 - 32.00 = -0.32$.

Example 4.16 Set up and implement a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price.

We are considering two scenarios: there is no difference or there is some difference in average prices.

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.
H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Can the t distribution be used for this application? The observations are based on a simple random sample from less than 10% of all books sold at the bookstore, so independence is reasonable. While the distribution is strongly skewed, the sample is reasonably large ($n = 73$), so we can proceed. Because the conditions are reasonably satisfied, we can apply the t distribution to this setting.

We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 14.26$) and the number of differences ($n_{diff} = 73$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{14.26}{\sqrt{73}} = 1.67$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, which is shown in Figure 4.14. The p-value is represented by the two (very) small tails.

To find the tail areas, we compute the test statistic, which is the T score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$T = \frac{\bar{x}_{diff} - 0}{SE_{x_{diff}}} = \frac{12.76 - 0}{1.67} = 7.59$$

The degrees of freedom are $df = 73 - 1 = 72$. If we examined Appendix C.2 on page 342, we would see that this value is larger than any in the 70 df row (we round down for df when using the table), meaning the two-tailed p-value is less than 0.01. If we used statistical software, we would find the p-value is less than 1-in-10 billion! Because the p-value is less than 0.05, we reject the null hypothesis. We have found convincing evidence that Amazon is, on average, cheaper than the UCLA bookstore for UCLA course textbooks.

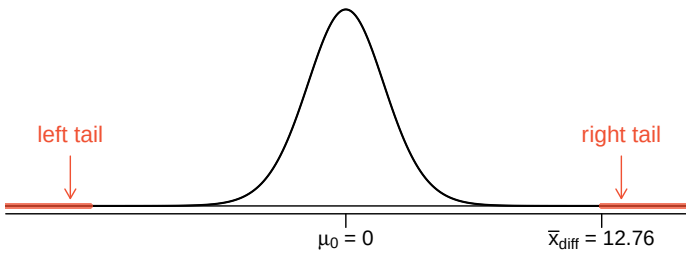


Figure 4.14: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

⊙ **Guided Practice 4.17** Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[14]

In the textbook price example, we applied the t distribution. However, as we mentioned in the last section, the t distribution looks a lot like the normal distribution when the degrees of freedom are larger than about 30. In such cases, including this one, it would be reasonable to use the normal distribution in place of the t distribution.

4.3 Difference of two means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. Just as with a single sample, we identify conditions to ensure we can use the t distribution with a point estimate of the difference, $\bar{x}_1 - \bar{x}_2$.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the impact of pregnant womens' smoking habits on birth weights of newborns, and exploring whether there is statistically significant evidence that one variations of an exam is harder than another variation. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

4.3.1 Confidence interval for a differences of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 4.15 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

A point estimate of the difference in the heart pumping variable can be found using the difference in the sample means:

$$\bar{x}_{esc} - \bar{x}_{control} = 3.50 - (-4.33) = 7.83$$

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Table 4.15: Summary statistics of the embryonic stem cell study.

[14] Conditions have already verified and the standard error computed in Example 4.16. To find the interval, identify $t^\star_{72}$ (use $df = 70$ in the table, $t^\star_{70} = 1.99$) and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 12.76 \pm 1.99 \times 1.67 \quad \rightarrow \quad (9.44, 16.08)$$

We are 95% confident that Amazon is, on average, between $9.44 and $16.08 cheaper than the UCLA bookstore for UCLA course books.

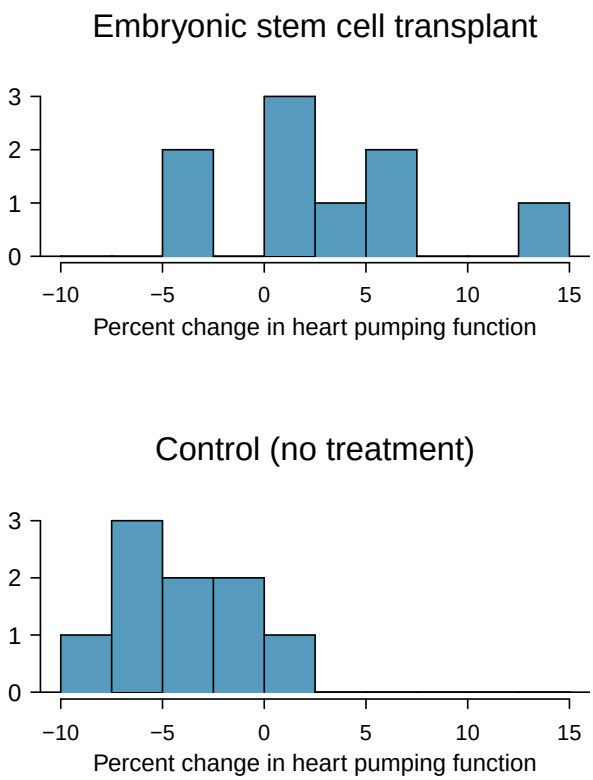


Figure 4.16: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

Using the t distribution for a difference in means

The t distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t distribution and (2) the samples are independent.

Example 4.18 Can the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$, be analyzed using the t distribution?

We check the two required conditions:

1. In this study, the sheep were independent of each other. Additionally, the distributions in Figure 4.16 don't show any clear deviations from normality, where we watch for prominent outliers in particular for such small samples. These findings imply each sample mean could itself be modeled using a t distribution.
2. The sheep in each group were also independent of each other.

Because both conditions are met, we can use the t distribution to model the difference of the two sample means.

Before we construct a confidence interval, we must calculate the standard error of the point estimate of the difference. For this, we use the following formula, where just as before we substitute the sample standard deviations into the formula:

$$SE_{\bar{x}_{esc}-\bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$
$$\approx \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Because we will use the t distribution, we also must identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will typically apply in the examples and guided practice.[15]

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t distribution and the standard error

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{4.19}$$

when each sample mean can itself be modeled using a t distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

Example 4.20 Calculate a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep after they've suffered a heart attack.

We will use the sample difference and the standard error for that point estimate from our earlier calculations:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83$$
$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Using $df = 8$, we can identify the appropriate $t^\star_{df} = t^\star_8$ for a 95% confidence interval as 2.31. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 7.83 \pm 2.31 \times 1.95 \quad \rightarrow \quad (3.38, 12.38)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.38% to 12.38%.

[15] This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method. In this example, computer software would have provided us a more precise degrees of freedom of $df = 12.225$.

4.3.2 Hypothesis tests based on a difference in means

A data set called `baby_smoke` represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table 4.17. 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? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 4.18.

	fAge	mAge	weeks	weight	sexBaby	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

Table 4.17: Four cases from the `baby_smoke` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

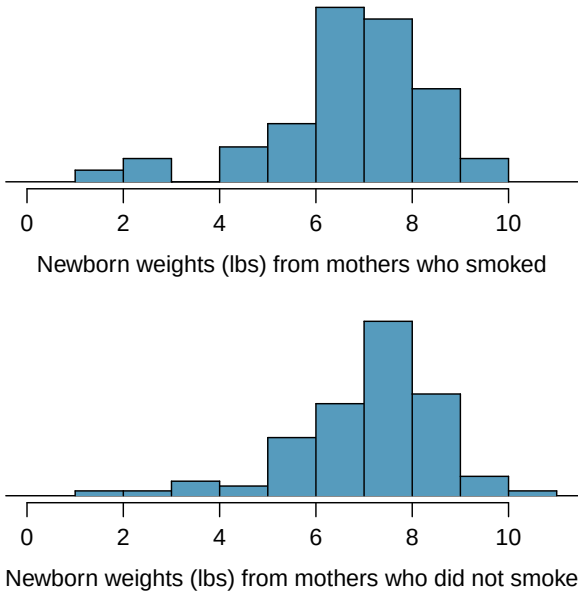


Figure 4.18: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

Example 4.21 Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

The null hypothesis represents the case of no difference between the groups.

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.

H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

We check the two conditions necessary to apply the t distribution to the difference in sample means. (1) Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, while each distribution is strongly skewed, the sample sizes of 50 and 100 would make it reasonable to model each mean separately using a t distribution. The skew is reasonable for these sample sizes of 50 and 100. (2) The independence reasoning applied in (1) also ensures the observations in each sample are independent. 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

Table 4.19: Summary statistics for the `baby_smoke` data set.

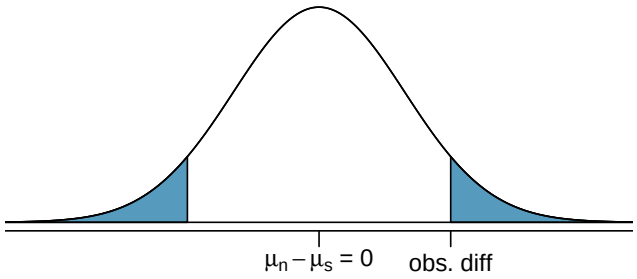
Guided Practice 4.22 The summary statistics in Table 4.19 may be useful for this exercise. (a) What is the point estimate of the population difference, $\mu_n - \mu_s$? (b) Compute the standard error of the point estimate from part (a).[16]

Example 4.23 Draw a picture to represent the p-value for the hypothesis test from Example 4.21.

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.

[16] (a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$. (b) The standard error of the estimate can be estimated using Equation (4.19):

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$



Example 4.24 Compute the p-value of the hypothesis test using the figure in Example 4.23, and evaluate the hypotheses using a significance level of $\alpha = 0.05$.

We start by computing the T score:

$$T = \frac{0.40 - 0}{0.26} = 1.54$$

Next, we compare this value to values in the t table in Appendix C.2 on page 342, where we use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The T score falls between the first and second columns in the $df = 49$ row of the t table, meaning the two-tailed p-value falls between 0.10 and 0.20 (reminder, find tail areas along the top of the table). This p-value is larger than the significance value, 0.05, so we fail to reject the null hypothesis. There is insufficient 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.

Guided Practice 4.25 Does the conclusion to Example 4.24 mean that smoking and average birth weight are unrelated?[17]

Guided Practice 4.26 If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect such a difference?[18]

4.3.3 Case study: two versions of a course exam

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 Table 4.20. 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.

Guided Practice 4.27 Construct a hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance.[19]

[17] Absolutely not. It is possible that there is some difference but we did not detect it. If there is a difference, we made a Type 2 Error. Notice: we also don't have enough information to, if there is an actual difference difference, confidently say which direction that difference would be in.

[18] 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.

[19] Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. H_A: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Table 4.20: Summary statistics of scores for each exam version.

⊙ **Guided Practice 4.28** To evaluate the hypotheses in Guided Practice 4.27 using the t distribution, we must first verify assumptions. (a) Does it seem reasonable that the scores are independent within each group? (b) What about the normality / skew condition for observations in each group? (c) Do you think scores from the two groups would be independent of each other, i.e. the two samples are independent?[20]

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated as

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62$$

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1$ and $n_2 - 1$: $df = 26$.

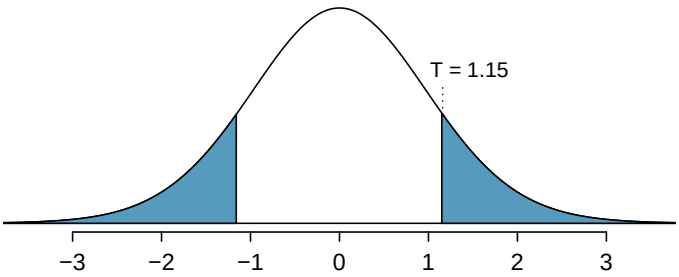


Figure 4.21: The t distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

[20] (a) It is probably reasonable to conclude the scores are independent, provided there was no cheating. (b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable. (c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.

Example 4.29 Identify the p-value using $df = 26$ and provide a conclusion in the context of the case study.

We examine row $df = 26$ in the t table. Because this value is smaller than the value in the left column, the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we do not reject the null hypothesis. That is, 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.

4.3.4 Summary for inference using the t distribution

Hypothesis tests. When applying the t distribution for a hypothesis test, we proceed as follows:

- Write appropriate hypotheses.
- Verify conditions for using the t distribution.
  - One-sample or differences from paired data: the observations (or differences) must be independent and nearly normal. For larger sample sizes, we can relax the nearly normal requirement, e.g. slight skew is okay or sample sizes of 15, moderate skew for sample sizes of 30, and strong skew for sample sizes of 60.
  - For a difference of means when the data are not paired: each sample mean must separately satisfy the one-sample conditions for the t distribution, and the data in the groups must also be independent.
- Compute the point estimate of interest, the standard error, and the degrees of freedom. For df, use $n - 1$ for one sample, and for two samples use either statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.
- Compute the T score and p-value.
- Make a conclusion based on the p-value, and write a conclusion in context and in plain language so anyone can understand the result.

Confidence intervals. Similarly, the following is how we generally computed a confidence interval using a t distribution:

- Verify conditions for using the t distribution. (See above.)
- Compute the point estimate of interest, the standard error, the degrees of freedom, and $t^\star_d f$.
- Calculate the confidence interval using the general formula, point estimate $\pm t^\star_{df} SE$.
- Put the conclusions in context and in plain language so even non-statisticians can understand the results.

4.3.5 Pooled standard deviation estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make the t distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s^2_{pooled} = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

Caution: Pooling standard deviations should be done only after careful research

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

4.4 Comparing many means with ANOVA (special topic)

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

Example 4.30 College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

The hypotheses may be written in the following form:

H_0: The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.

H_A: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

Example 4.31 Examine Figure 4.22. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

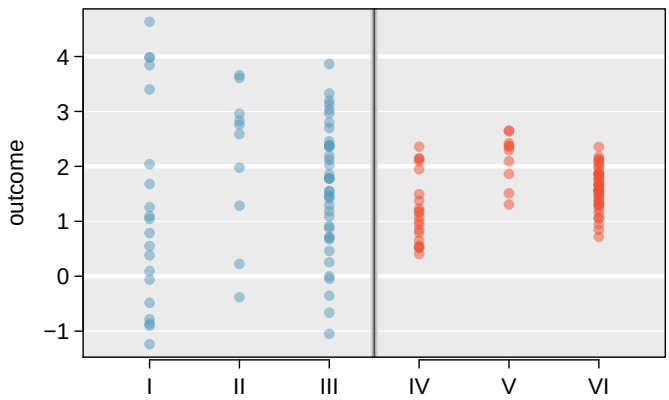


Figure 4.22: Side-by-side dot plot for the outcomes for six groups.

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large *relative to the variability in the individual observations within each group.*

4.4.1 Is batting performance related to player position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: outfielder (OF), infielder (IF), designated hitter (DH), and catcher (C). We will use a data set called bat10, which includes batting records of 327 Major League Baseball (MLB) players from the 2010 season. Six of the 327 cases represented in bat10 are shown in Table 4.23, and descriptions for each variable are provided in Table 4.24. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (OBP). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NYM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

Table 4.23: Six cases from the bat10 data matrix.

variable	description
name	Player name
team	The abbreviated name of the player's team
position	The player's primary field position (OF, IF, DH, C)
AB	Number of opportunities at bat
H	Number of hits
HR	Number of home runs
RBI	Number of runs batted in
AVG	Batting average, which is equal to H/AB
OBP	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Table 4.24: Variables and their descriptions for the bat10 data set.

Guided Practice 4.32 The null hypothesis under consideration is the following: $\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_{C}$. Write the null and corresponding alternative hypotheses in plain language.[21]

Example 4.33 The player positions have been divided into four groups: outfield (OF), infield (IF), designated hitter (DH), and catcher (C). What would be an appropriate point estimate of the on-base percentage by outfielders, μ_{OF}?

A good estimate of the on-base percentage by outfielders would be the sample average of AVG for just those players whose position is outfield: $\bar{x}_{OF} = 0.334$.

[21] H_0: The average on-base percentage is equal across the four positions. H_A: The average on-base percentage varies across some (or all) groups.

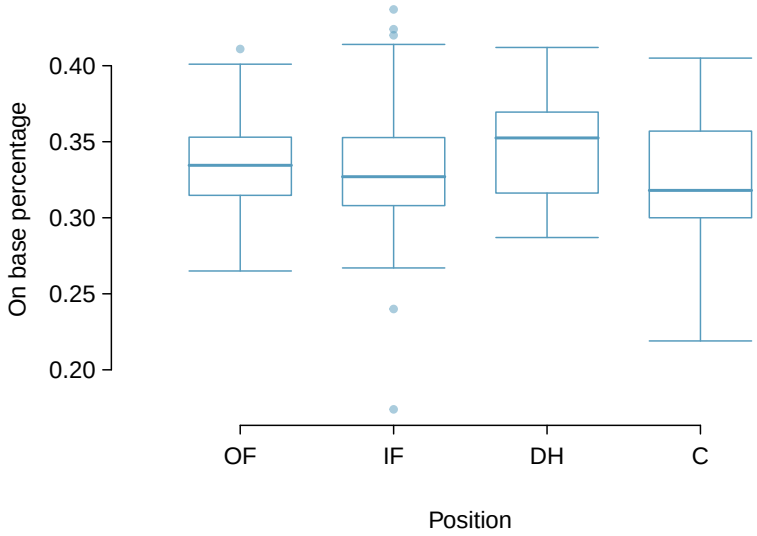


Figure 4.26: Side-by-side box plot of the on-base percentage for 327 players across four groups. There is one prominent outlier visible in the infield group, but with 154 observations in the infield group, this outlier is not a concern.

Table 4.25 provides summary statistics for each group. A side-by-side box plot for the on-base percentage is shown in Figure 4.26. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean ($\bar{x}_i$)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045

Table 4.25: Summary statistics of on-base percentage, split by player position.

Example 4.34 The largest difference between the sample means is between the designated hitter and the catcher positions. Consider again the original hypotheses:

H_0: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{DH}} = \mu_{\texttt{C}}$
H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of $\mu_{\texttt{DH}}$ and $\mu_{\texttt{C}}$ is statistically significant at a 0.05 significance level?

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large

differences for the formal test, leading to an inflation in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional information on the ideas expressed in Example 4.34, we recommend reading about the **prosecutor's fallacy**.[22]

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in means could have happened just by chance even if there was no difference in the respective population means.

4.4.2 Analysis of variance (ANOVA) and the F test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups (MSG)**, and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote,[23] however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error (MSE)**, which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE are provided in the footnote[24] for interested readers.

[22] See, for example, www.stat.columbia.edu/∼cook/movabletype/archives/2007/05/the_prosecutors.html.

[23] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i.

[24] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total (SST)** is computed as

$$SST = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

where the sum is over all observations in the data set. Then we compute the **sum of squared errors**

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE} \tag{4.35}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

⊙ **Guided Practice 4.36** For the baseball data, $MSG = 0.00252$ and $MSE = 0.00127$. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 1.994.[25]

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 3 and 323 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 4.27.

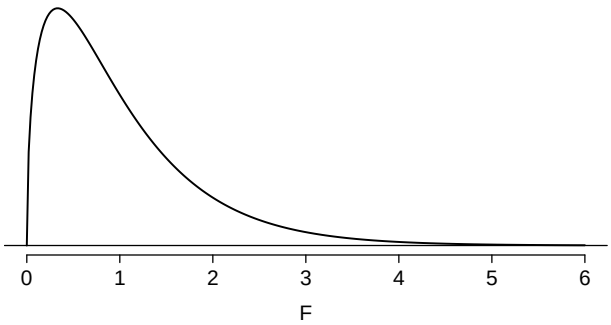


Figure 4.27: An F distribution with $df_1 = 3$ and $df_2 = 323$.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

(SSE) in one of two equivalent ways:

$$\begin{aligned} SSE &= SST - SSG \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \end{aligned}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

[25] There are $k = 4$ groups, so $df_G = k - 1 = 3$. There are $n = n_1 + n_2 + n_3 + n_4 = 327$ total observations, so $df_E = n - k = 323$. Then the F statistic is computed as the ratio of MSG and MSE: $F = \frac{MSG}{MSE} = \frac{0.00252}{0.00127} = 1.984 \approx 1.994$. ($F = 1.994$ was computed by using values for MSG and MSE that were not rounded.)

The F statistic and the F test

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

⊙ **Guided Practice 4.37** The test statistic for the baseball example is $F = 1.994$. Shade the area corresponding to the p-value in Figure 4.27. [26]

● **Example 4.38** The p-value corresponding to the shaded area in the solution of Guided Practice 4.37 is equal to about 0.115. Does this provide strong evidence against the null hypothesis?

The p-value is larger than 0.05, indicating the evidence is not strong enough to reject the null hypothesis at a significance level of 0.05. That is, the data do not provide strong evidence that the average on-base percentage varies by player's primary field position.

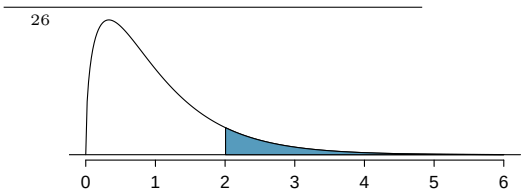
4.4.3 Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we will see in Chapters 5 and 6. Table 4.28 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F test statistic and p-value can be retrieved from the last columns.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	3	0.0076	0.0025	1.9943	0.1147
Residuals	323	0.4080	0.0013		
			$s_{pooled} = 0.036$ on $df = 323$		

Table 4.28: ANOVA summary for testing whether the average on-base percentage differs across player positions.



4.4.4 Graphical diagnostics for an ANOVA analysis

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

Independence. If the data are a simple random sample from less than 10% of the population, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). For example, in the MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations.

Approximately normal. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small. The normal probability plots for each group of the MLB data are shown in Figure 4.29; there is some deviation from normality for infielders, but this isn't a substantial concern since there are about 150 observations in that group and the outliers are not extreme. Sometimes in ANOVA there are so many groups or so few observations per group that checking normality for each group isn't reasonable. See the footnote[27] for guidance on how to handle such instances.

Constant variance. The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a side-by-side box plot of the outcomes across the groups, as in Figure 4.26 on page 187. In this case, the variability is similar in the four groups but not identical. We see in Table 4.25 on page 187 that the standard deviation varies a bit from one group to the next. Whether these differences are from natural variation is unclear, so we should report this uncertainty with the final results.

Caution: Diagnostics for an ANOVA analysis

Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

4.4.5 Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t test, but we use a modified significance level and a pooled estimate of the standard deviation across groups. Usually this pooled standard deviation can be found in the ANOVA table, e.g. along the bottom of Table 4.28.

[27] First calculate the **residuals** of the baseball data, which are calculated by taking the observed values and subtracting the corresponding group means. For example, an outfielder with OBP of 0.435 would have a residual of $0.405 - \bar{x}_{OF} = 0.071$. Then to check the normality condition, create a normal probability plot using all the residuals simultaneously.

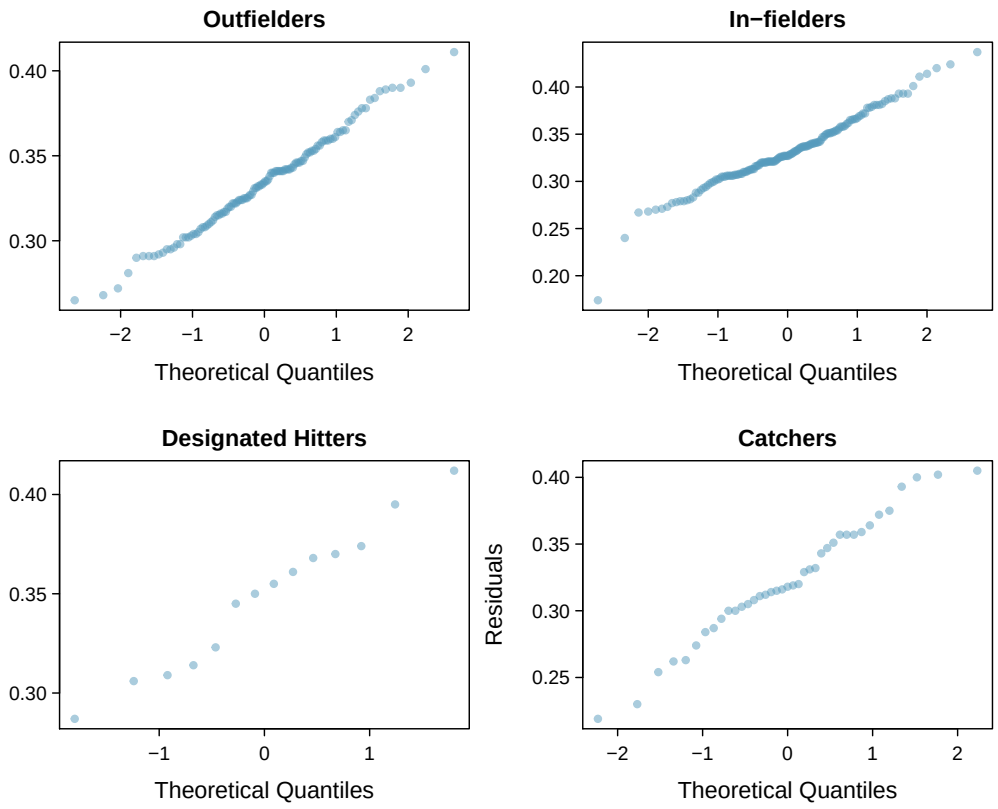


Figure 4.29: Normal probability plot of OBP for each field position.

Example 4.39 Example 4.30 on page 185 discussed three statistics lectures, all taught during the same semester. Table 4.30 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 4.31. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. No such situations were evident for these particular data, and we believe that independence is acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be verified in Table 4.30, supporting the constant variance assumption.

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.1	72.0	78.9
s_i	13.9	13.8	13.1

Table 4.30: Summary statistics for the first midterm scores in three different lectures of the same course.

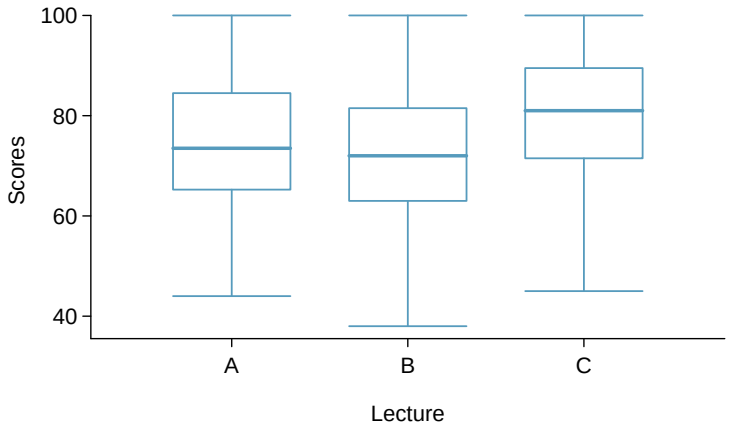


Figure 4.31: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

⊙ **Guided Practice 4.40** An ANOVA was conducted for the midterm data, and summary results are shown in Table 4.32. What should we conclude?[28]

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 4.34 on page 187: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modified significance level.

[28]The p-value of the test is 0.0330, less than the default significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the difference in the average midterm scores are not due to chance.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		
			$s_{pooled} = 13.61$ on $df = 161$		

Table 4.32: ANOVA summary table for the midterm data.

Multiple comparisons and the Bonferroni correction for α

The scenario of testing many pairs of groups is called **multiple comparisons**. The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha / K$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Example 4.41 In Guided Practice 4.40, you found strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modified significance level of $\alpha^* = 0.05/3 = 0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled} = 13.61$ on $df = 161$, which is provided in the ANOVA summary table.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{x}_A - \bar{x}_B = 75.1 - 72 = 3.1 \qquad SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56$$

(See Section 4.3.5 on page 183 for additional details.) This results in a T score of 1.21 on $df = 161$ (we use the df associated with s_{pooled}). Statistical software was used to precisely identify the two-tailed p-value since the modified significance of 0.0167 is not found in the t table. The p-value (0.228) is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on $df = 161$ and a two-tailed p-value of 0.1462. This p-value is larger than α^*, so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on $df = 161$ and a two-tailed p-value of 0.0102. This p-value is smaller than α^*. Here we find strong evidence of a difference in the means of lectures B and C.

We might summarize the findings of the analysis from Example 4.41 using the following notation:

$$\mu_A \stackrel{?}{=} \mu_B \qquad \mu_A \stackrel{?}{=} \mu_C \qquad \mu_B \neq \mu_C$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

Caution: Sometimes an ANOVA will reject the null but no groups will have statistically significant differences

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion.* It only means we have not been able to successfully identify which groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

4.5 Bootstrapping to study the standard deviation

We analyzed textbook pricing data in Section 4.2 and found that prices on Amazon were statistically significantly cheaper on average. We might also want to better understand the variability of the price difference from one book to another, which we quantified using the standard deviation: $s = \$14.26$. The sample standard deviation is a point estimate for the population standard deviation. Just as we care about the precision of a sample mean, we may care about the precise of the sample standard deviation.

4.5.1 Bootstrap samples and distributions

The theory required to quantify the uncertainty of the sample standard deviation is complex. In an ideal world, we would sample data from the population again and recompute the standard deviation with this new sample. Then we could do it again. And again. And so on until we get enough standard deviation estimates that we have a good sense of the precision of our original estimate. This is an ideal world where sampling data is free or extremely cheap. That is rarely the case, which poses a challenge to this "resample from the population" approach.

However, we can sample from the sample. In the textbook pricing example, there are 73 price differences. This sample can serve as a proxy for the population: we sample from this data set to get a sense for what it would be like if we took new samples.

A **bootstrap sample** is a sample of the original sample. In the case of the textbook data, we proceed as follows:

1. Randomly sample one observation from the 73 price differences.
2. Randomly sample a second observation from the 73 price differences. There is a 1-in-73 chance that this second observation will be the same one sampled in the first step.

   ⋮

73. Randomly sample a 73^{rd} observation from the 73 price differences.

This type of sampling is called **sampling with replacement**. Table 4.33 shows a bootstrap sample for the textbook pricing example. Some of the values, such as **16.80**, are duplicated since occasionally we sample the same observation multiple times.

16.80	6.63	5.39	6.39	14.05	6.63	-0.25	12.45	-0.22	9.45	9.45
11.70	39.08	4.80	28.72	9.45	-0.25	-3.88	2.82	45.34	28.72	16.62
38.35	4.74	44.40	3.74	1.75	2.84	30.25	3.35	6.63	30.50	0.00
4.96	6.39	9.48	**16.80**	66.00	44.40	-0.25	-2.55	17.98	2.82	
29.29	9.22	11.70	9.31	4.80	13.63	9.45	38.23	4.96	19.69	
14.26	12.45	5.39	-0.28	8.23	0.42	2.82	4.78	7.01	4.64	
9.12	9.31	9.12	11.70	27.15	28.72	30.71	2.84	-9.53	14.05	

Table 4.33: A bootstrap sample of the textbook price differences, which represents a sample of 73 values from the original 73 observations, where we are sampling with replacement. In sampling with replacement, it is possible for a value to be sampled multiple times. For example, **16.80** was sampled twice in this bootstrap sample.

A bootstrap sample behaves similarly to how an actual sample would behave, and we compute the point estimate of interest. In the textbook price example, we compute the standard deviation of the bootstrap sample: $13.98.

4.5.2 Inference using the bootstrap

One bootstrap sample is not enough to understand the uncertainty of the standard deviation, so we need to collect another bootstrap sample and compute the standard deviation: $16.21. And another: $14.07. And so on. Using a computer, we took 10,000 bootstrap samples and computed the standard deviation for each, and these are summarized in Figure 4.34. This is called the **bootstrap distribution** of the standard deviation for the textbook price differences. To make use of this distribution, we make an important assumption: the bootstrap distribution shown in Figure 4.34 is similar to the sampling distribution of the standard deviation. This assumption is reasonable when doing an informal exploration of the uncertainty of an estimate, and under certain conditions, we can rely on it for more formal inference methods.

Example 4.42 Describe the bootstrap distribution for the standard deviation shown in Figure 4.34.

The distribution is symmetric, bell-shaped, and centered near $14.26, which is the point estimate from the original data. The standard deviation of the bootstrap distribution is $1.60, and most observations in this distribution lie between $11 and $17.

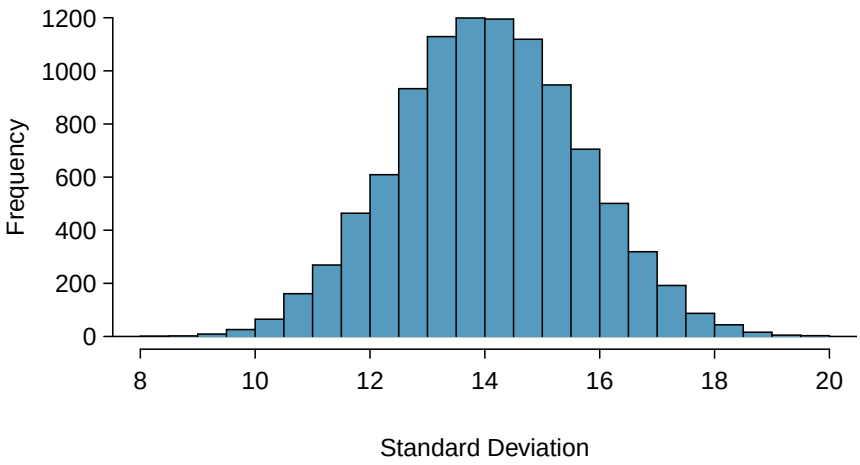


Figure 4.34: Bootstrap distribution for the standard deviation of textbook price differences. The distribution is approximately centered at the original sample's standard deviation, \$14.26.

In this example, the bootstrap distribution's standard deviation, \$1.60, quantifies the uncertainty of the point estimate. This is an estimate of the standard error based on the bootstrap. We might be tempted to use it for a 95% confidence interval, but first we must perform some due diligence. As with every statistical method, we must check certain conditions before performing formal inference using the bootstrap.

Bootstrapping for the Standard Deviation

The bootstrap distribution for the standard deviation will be a good approximation of the sampling distribution for the standard deviation when

1. observations in the original sample are independent,
2. the original sample size is at least 30, and
3. the bootstrap distribution is nearly normal.

We're already familiar with checking independence of observations, which we previously checked for this data set, and the second condition is easy to check. The last condition can be checked by examining the bootstrap distribution using a normal probability plot, as shown in Figure 4.35. In this example, we see a very straight line, which indicates the bootstrap distribution is nearly normal, and we can move forward with constructing a confidence interval.

As with many other point estimates, we will use the familiar formula

$$\text{point estimate} \pm t^\star_{df} \times SE$$

In the textbook example, using $df = 73 - 1 = 72$ leads to $t^\star_{72} = 1.99$ for a 95% confidence level. For bootstrapping, the standard error is computed as the standard deviation of the bootstrap distribution.

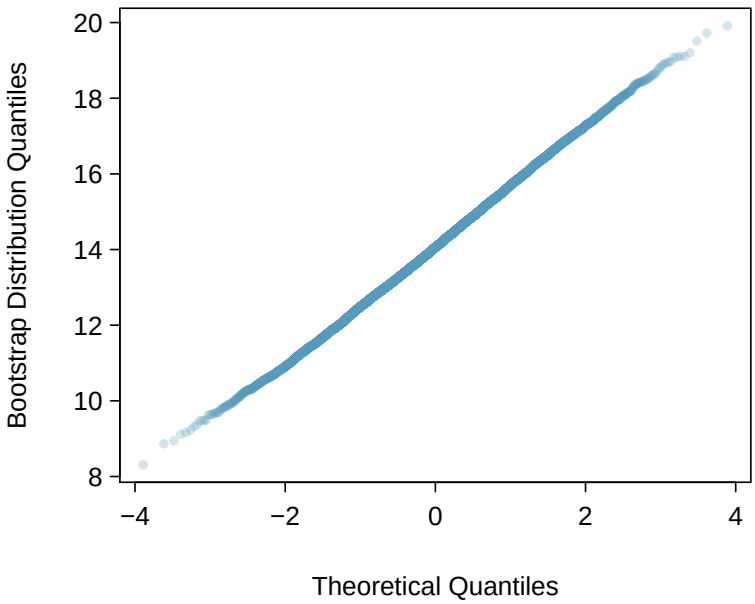


Figure 4.35: Normal probability plot for the bootstrap distribution.

Example 4.43 Compute the 95% confidence interval for the standard deviation of the textbook price difference.

We use the general formula for a 95% confidence interval with the t distribution:

$$\begin{aligned} &\text{point estimate} \pm t^\star_{df} \times SE \\ &14.26 \pm 1.99 \times 1.60 \\ &(\$11.08, \$17.44) \end{aligned}$$

We are 95% confident that the standard deviation of the textbook price differences is between \$11.08 and \$17.44.

Had we wanted to conduct a hypothesis test, we could have used the point estimate and standard error for a t test as we have in previous sections.

Bootstrap for other parameters
The bootstrap may be used with any parameters using the same conditions as were provided for the standard deviation. However, in other situations, it may be more important to examine the validity of the third condition: that the bootstrap distribution is nearly normal.

4.5.3 Frequently asked questions

There are more types of bootstrap techniques, right? Yes! There are many excellent bootstrap techniques. We have only chosen to present one bootstrap technique that could be explained in a single section and is also reasonably reliable.

Can we use the bootstrap for the mean or difference of means? Technically, yes. However, the methods introduced earlier tend to be more reliable than this particular bootstrapping method and other simple bootstrapping techniques. See the following page for details on an investigation into the accuracy of several bootstrapping methods as well as the t distribution method introduced earlier in this chapter:

www.openintro.org/stat/bootstrap

I've heard a technique called the percentile bootstrap that is very robust. It is a commonly held belief that the percentile bootstrap is a robust bootstrap method. That is false. The percentile method is one of the least reliable bootstrap methods. Instead, use the method described in this section, which is more reliable, or learn about more advanced techniques.

4.6 Exercises

4.6.1 One-sample means with the t distribution

4.1 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%

4.2 Working backwards, Part I. 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.

4.3 Working backwards, Part II. 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.

4.4 Find the p-value. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given set of hypotheses and T test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

(a) $H_A : \mu > \mu_0$, $n = 11$, $T = 1.91$
(b) $H_A : \mu < \mu_0$, $n = 17$, $T = -3.45$
(c) $H_A : \mu \neq \mu_0$, $n = 7$, $T = 0.83$
(d) $H_A : \mu > \mu_0$, $n = 28$, $T = 2.13$

4.5 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. Do these data provide strong evidence that New Yorkers sleep more or less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(a) Write the hypotheses in symbols and in words.

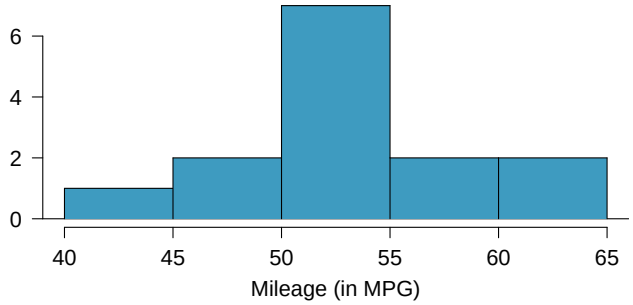
(b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

(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 95% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

4.6 Fuel efficiency of Prius. Fueleconomy.gov, the official US government source for fuel economy information, allows users to share gas mileage information on their vehicles. The histogram below shows the distribution of gas mileage in miles per gallon (MPG) from 14 users who drive a 2012 Toyota Prius. The sample mean is 53.3 MPG and the standard deviation is 5.2 MPG. Note that these data are user estimates and since the source data cannot be verified, the accuracy of these estimates are not guaranteed.[29]



(a) We would like to use these data to evaluate the average gas mileage of all 2012 Prius drivers. Do you think this is reasonable? Why or why not?

(b) The EPA claims that a 2012 Prius gets 50 MPG (city and highway mileage combined). Do these data provide strong evidence against this estimate for drivers who participate on fueleconomy.gov? Note any assumptions you must make as you proceed with the test.

(c) Calculate a 95% confidence interval for the average gas mileage of a 2012 Prius by drivers who participate on fueleconomy.gov.

4.7 Find the mean. You are given the following hypotheses:

$$H_0 : \mu = 60$$
$$H_A : \mu < 60$$

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.

4.8 $t^\star$ 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.

[29] Fuelecomy.gov, Shared MPG Estimates: Toyota Prius 2012.

4.6.2 Paired data

4.9 Climate change, Part I. Is there strong evidence of climate change? Let's consider a small scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. Then the difference between the two readings (temperature in 2008 - temperature in 1968) was calculated for each of the 51 different locations. The average of these 51 values was 1.1 degrees with a standard deviation of 4.9 degrees.

(a) Is there a relationship between the observations collected in 1968 and 2008? 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.

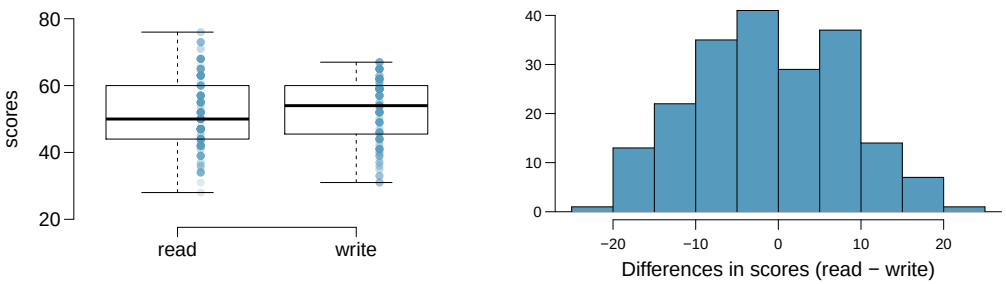
(d) Calculate the test statistic and find the p-value.

(e) What do you conclude? Interpret your conclusion in context.

(f) What type of error might we have made? Explain in context what the error means.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the temperature measurements from 1968 and 2008 to include 0? Explain your reasoning.

4.10 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.

4.11 Climate change, Part II. We considered the differences between the temperature readings in January 1 of 1968 and 2008 at 51 locations in the continental US in Exercise 4.9. The mean and standard deviation of the reported differences are 1.1 degrees and 4.9 degrees.

(a) Calculate a 95% confidence interval for the average difference between the temperature measurements between 1968 and 2008.

(b) Interpret this interval in context.

(c) Does the confidence interval provide convincing evidence that the temperature was different in 2008 than in 1968 in the continental US? Explain.

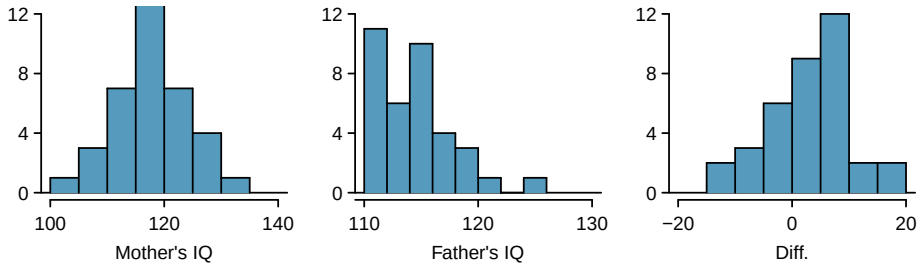
4.12 High school and beyond, Part II. We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey in Exercise 4.11. The mean and standard deviation of the differences are $\bar{x}_{read-write} = -0.545$ and 8.887 points.

(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.

4.13 Gifted children. Researchers collected a simple random sample of 36 children who had been identified as gifted in a large city. The following histograms show the distributions of the IQ scores of mothers and fathers of these children. Also provided are some sample statistics.[30]



	Mother	Father	Diff.
Mean	118.2	114.8	3.4
SD	6.5	3.5	7.5
n	36	36	36

(a) Are the IQs of mothers and the IQs of fathers in this data set related? Explain.

(b) Conduct a hypothesis test to evaluate if the scores are equal on average. Make sure to clearly state your hypotheses, check the relevant conditions, and state your conclusion in the context of the data.

4.14 Paired or not? 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 for Intel's stock and another random sample of 50 days for Southwest's stock.

(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.

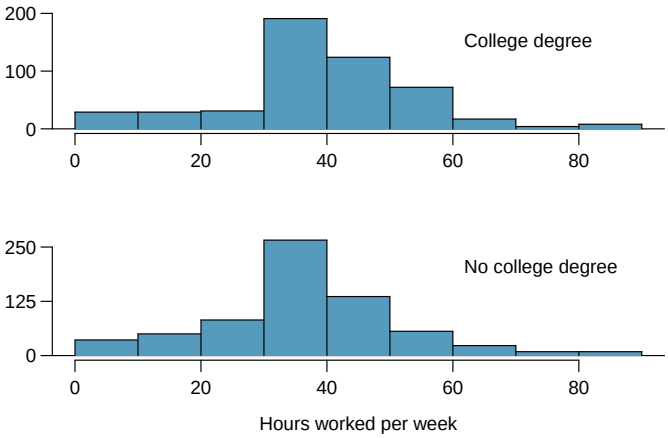
(c) A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

[30] F.A. Graybill and H.K. Iyer. *Regression Analysis: Concepts and Applications*. Duxbury Press, 1994, pp. 511–516.

4.6.3 Difference of two means

4.15 Math scores of 13 year olds, Part I. The National Assessment of Educational Progress tested a simple random sample of 1,000 thirteen year old students in both 2004 and 2008 (two separate simple random samples). The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively. Calculate a 90% confidence interval for the change in average scores from 2004 to 2008, and interpret this interval in the context of the application. (Reminder: check conditions.)[31]

4.16 Work hours and education, Part I. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. The histograms below display the distributions of hours worked per week for two education groups: those with and without a college degree.[32] Suppose we want to estimate the average difference between the number of hours worked per week by all Americans with a college degree and those without a college degree. Summary information for each group is shown in the tables.



	College degree
Mean	41.8 hrs
SD	15.1 hrs
n	505

	No college degree
Mean	39.4 hrs
SD	15.1 hrs
n	667

(a) What is the parameter of interest, and what is the point estimate?

(b) Are conditions satisfied for estimating this difference using a confidence interval?

(c) Create a 95% confidence interval for the difference in number of hours worked between the two groups, and interpret the interval in context.

(d) Can you think of any real world justification for your results? (*Note:* There isn't a single correct answer to this question.)

4.17 Math scores of 13 year olds, Part II. Exercise 4.15 provides data on the average math scores from tests conducted by the National Assessment of Educational Progress in 2004 and 2008. Two separate simple random samples were taken in each of these years. The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively.

(a) Do these data provide strong evidence that the average math score for 13 year old students has changed from 2004 to 2008? Use a 10% significance level.

(b) It is possible that your conclusion in part (a) is incorrect. What type of error is possible for this conclusion? Explain.

(c) Based on your hypothesis test, would you expect a 90% confidence interval to contain the null value? Explain.

[31] National Center for Education Statistics, NAEP Data Explorer.
[32] National Opinion Research Center, General Social Survey, 2010.

4.18 Work hours and education, Part II. The General Social Survey described in Exercise 4.16 included random samples from two groups: US residents with a college degree and US residents without a college degree. For the 505 sampled US residents with a college degree, the average number of hours worked each week was 41.8 hours with a standard deviation of 15.1 hours. For those 667 without a degree, the mean was 39.4 hours with a standard deviation of 15.1 hours. Conduct a hypothesis test to check for a difference in the average number of hours worked for the two groups.

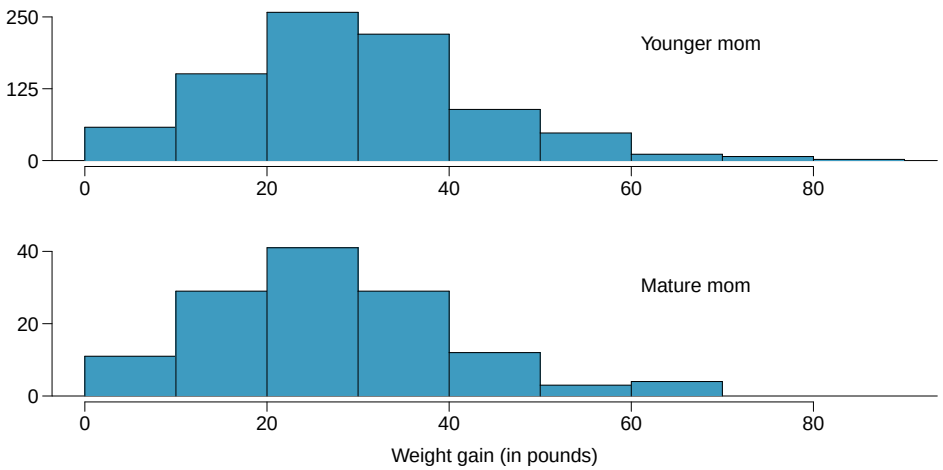
4.19 Does the Paleo diet work? The Paleo diet allows only for foods that humans typically consumed over the last 2.5 million years, excluding those agriculture-type foods that arose during the last 10,000 years or so. Researchers randomly divided 500 volunteers into two equal-sized groups. One group spent 6 months on the Paleo diet. The other group received a pamphlet about controlling portion sizes. Randomized treatment assignment was performed, and at the beginning of the study, the average difference in weights between the two groups was about 0. After the study, the Paleo group had lost on average 7 pounds with a standard deviation of 20 pounds while the control group had lost on average 5 pounds with a standard deviation of 12 pounds.

(a) The 95% confidence interval for the difference between the two population parameters (Paleo - control) is given as (-0.891, 4.891). Interpret this interval in the context of the data.

(b) Based on this confidence interval, do the data provide convincing evidence that the Paleo diet is more effective for weight loss than the pamphlet (control)? Explain your reasoning.

(c) Without explicitly performing the hypothesis test, do you think that if the Paleo group had lost 8 instead of 7 pounds on average, and everything else was the same, the results would then indicate a significant difference between the treatment and control groups? Explain your reasoning.

4.20 Weight gain during pregnancy. In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relationship between habits and practices of expectant mothers and the birth of their children. The following histograms show the distributions of weight gain during pregnancy by 867 younger moms (less than 35 years old) and 133 mature moms (35 years old and over) who have been randomly sampled from this large data set. The average weight gain of younger moms is 30.56 pounds, with a standard deviation of 14.35 pounds, and the average weight gain of mature moms is 28.79 pounds, with a standard deviation of 13.48 pounds. Calculate a 95% confidence interval for the difference between the average weight gain of younger and mature moms. Also comment on whether or not this interval provides strong evidence that there is a significant difference between the two population means.

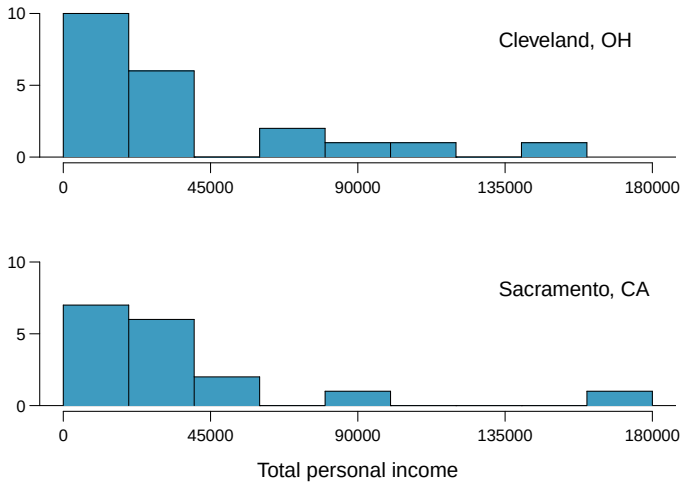


4.21 Body fat in women and men. The third National Health and Nutrition Examination Survey collected body fat percentage (BF) data from 13,601 subjects whose ages are 20 to 80. A summary table for these data is given below. Note that BF is given as *mean* ± *standard error*. Construct a 95% confidence interval for the difference in average body fat percentages between men and women, and explain the meaning of this interval. Tip: the standard error can be calculated as $SE = \sqrt{SE_M^2 + SE_W^2}$.[33]

Gender	n	BF (%)
Men	6,580	23.9 ± 0.07
Women	7,021	35.0 ± 0.09

4.22 Child care hours, Part I. The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments. One of the variables collected on the survey is the number of hours parents spend taking care of children in their household under age 6 (feeding, bathing, dressing, holding, or watching them). In 2006, 487 females and 312 males were surveyed for this question. On average, females reported spending 31 hours with a standard deviation of 31 hours, and males reported spending 16 hours with a standard deviation of 21 hours. Calculate a 95% confidence interval for the difference between the average number of hours Chinese males and females spend taking care of their children under age 6. Also comment on whether this interval suggests a significant difference between the two population parameters. You may assume that conditions for inference are satisfied.[34]

4.23 Cleveland vs. Sacramento. Average income varies from one region of the country to another, and it often reflects both lifestyles and regional living expenses. Suppose a new graduate is considering a job in two locations, Cleveland, OH and Sacramento, CA, and he wants to see whether the average income in one of these cities is higher than the other. He would like to conduct a t test based on two small samples from the 2000 Census, but he first must consider whether the conditions are met to implement the test. Below are histograms for each city. Should he move forward with the t test? Explain your reasoning.



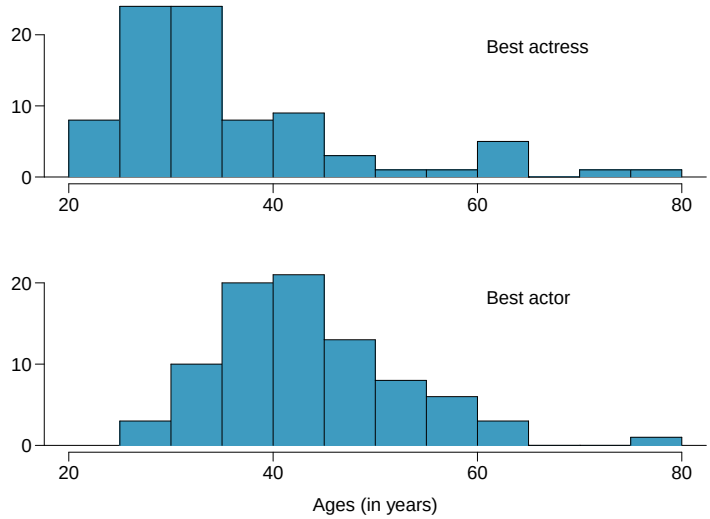
	Cleveland, OH
Mean	$ 35,749
SD	$ 39,421
n	21

	Sacramento, CA
Mean	$ 35,500
SD	$ 41,512
n	17

[33] A Romero-Corral et al. "Accuracy of body mass index in diagnosing obesity in the adult general population". In: *International Journal of Obesity* 32.6 (2008), pp. 959–966.

[34] UNC Carolina Population Center, China Health and Nutrition Survey, 2006.

4.24 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 2012. Summary statistics for these distributions are also provided. Is a t test appropriate for evaluating whether the difference in the average ages of best actors and actresses might be due to chance? Explain your reasoning.[35]

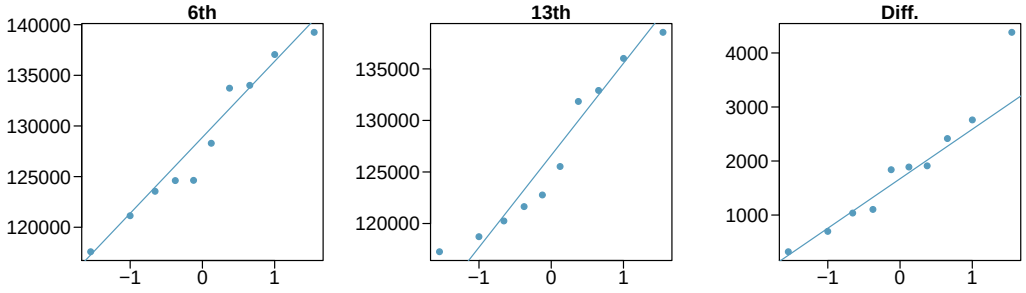


	Best Actress
Mean	35.6
SD	11.3
n	84

	Best Actor
Mean	44.7
SD	8.9
n	84

[35] Oscar winners from 1929 – 2012, data up to 2009 from the Journal of Statistics Education data archive and more current data from wikipedia.org.

4.25 Friday the 13th, 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 13th and the previous Friday, Friday the 6th. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6th and Friday the 13th 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.[36]



	6th	13th	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 6th is different than the number out on Friday the 13th?

(c) Check conditions to carry out the hypothesis test from part (b).

(d) Calculate the test statistic and the p-value.

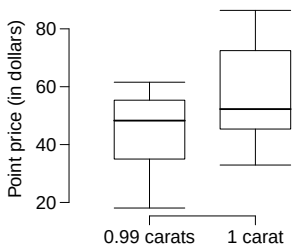
(e) What is the conclusion of the hypothesis test?

(f) Interpret the p-value in this context.

(g) What type of error might have been made in the conclusion of your test? Explain.

4.26 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.[37]

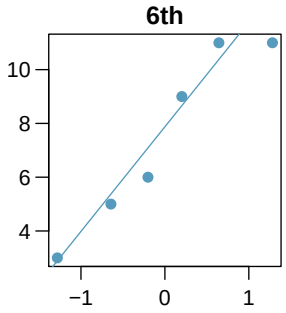
Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.

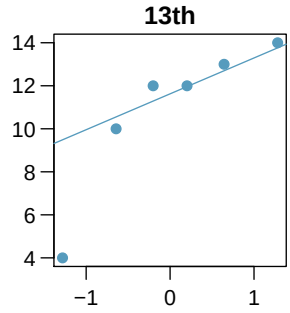


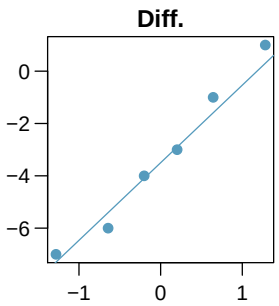
	0.99 carats	1 carat
Mean	$ 44.51	$ 56.81
SD	$ 13.32	$ 16.13
n	23	23

[36] T.J. Scanlon et al. "Is Friday the 13th Bad For Your Health?" In: *BMJ* 307 (1993), pp. 1584–1586.
[37] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

4.27 Friday the 13th, Part II. The Friday the 13^{th} study reported in Exercise 4.25 also provides data on traffic accident related emergency room admissions. The distributions of these counts from Friday the 6^{th} and Friday the 13^{th} are shown below for six such paired dates along with summary statistics. You may assume that conditions for inference are met.







	6^{th}	13^{th}	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 6^{th} and Friday the 13^{th}.

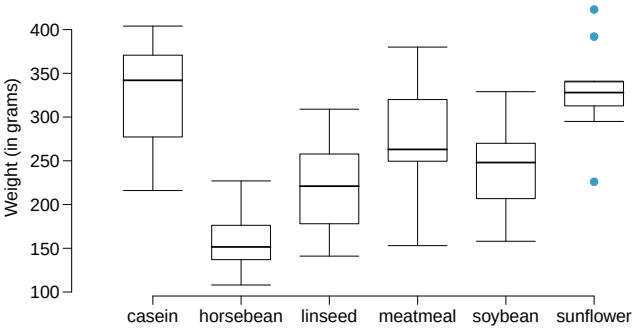
(b) Calculate a 95% confidence interval for the difference between the average numbers of traffic accident related emergency room admissions between Friday the 6^{th} and Friday the 13^{th}.

(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.

4.28 Diamonds, Part II. In Exercise 4.26, 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

4.29 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.[38]



	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.

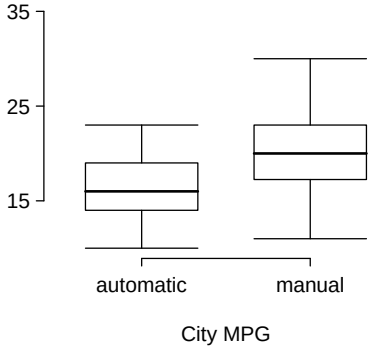
(b) Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different? Use a 5% significance level.

(c) What type of error might we have committed? Explain.

(d) Would your conclusion change if we used $\alpha = 0.01$?

4.30 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 manufactured in 2012. 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.[39]

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26

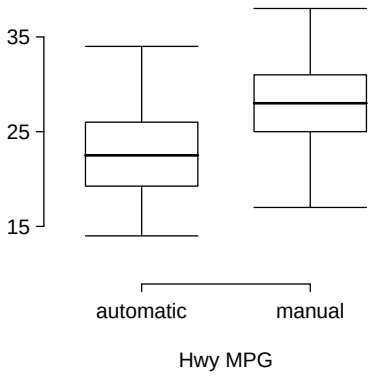


4.31 Chicken diet and weight, Part II. Casein is a common weight gain supplement for humans. Does it have an effect on chickens? Using data provided in Exercise 4.29, test the hypothesis that the average weight of chickens that were fed casein is different than the average weight of chickens that were 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.

[38] Chicken Weights by Feed Type, from the `datasets` package in R..
[39] U.S. Department of Energy, Fuel Economy Data, 2012 Datafile.

4.32 Fuel efficiency of manual and automatic cars, Part II. The table provides summary statistics on highway fuel economy of cars manufactured in 2012 (from Exercise 4.30). 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.[40]

	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26



4.33 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.[41]

4.34 Gaming and distracted eating, Part II. The researchers from Exercise 4.33 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?

[40] U.S. Department of Energy, Fuel Economy Data, 2012 Datafile.

[41] 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.

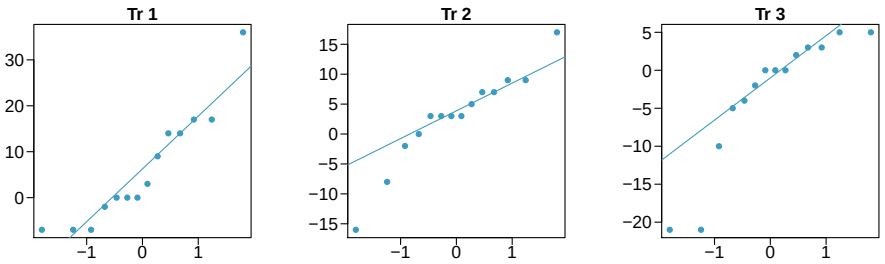
4.35 Prison isolation experiment, Part I. Subjects from Central Prison in Raleigh, NC, volunteered for an experiment involving an "isolation" experience. The goal of the experiment was to find a treatment that reduces subjects' psychopathic deviant T scores. This score measures a person's need for control 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.[42]



	Tr 1	Tr 2	Tr 3
Mean	6.21	2.86	-3.21
SD	12.3	7.94	8.57
n	14	14	14

4.36 True or false, Part I. 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.

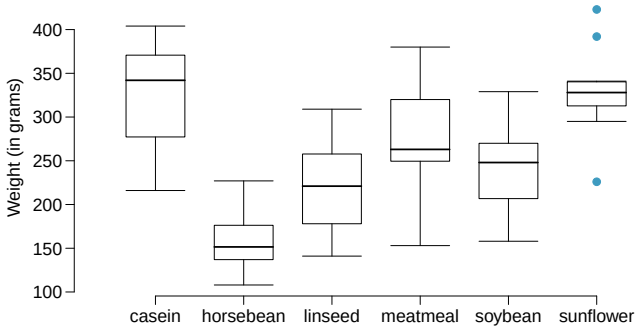
[42] Prison isolation experiment.

4.6.4 Comparing many means with ANOVA

4.37 Chicken diet and weight, Part III. In Exercises 4.29 and 4.31 we compared the effects of two types of feed at a time. A better analysis would first consider all feed types at once: casein, horsebean, linseed, meat meal, soybean, and sunflower. The ANOVA output below can be used to test for differences between the average weights of chicks on different diets.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231,129.16	46,225.83	15.36	0.0000
Residuals	65	195,556.02	3,008.55		

Conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions. Figures and summary statistics are shown below.



	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

4.38 Student performance across discussion sections. A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
$\bar{x}_i$	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
section	7	525.01	75.00	1.87	0.0767
Residuals	189	7584.11	40.13		

Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to proceed with the test.

4.39 Coffee, depression, and physical activity. Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.[43]

	Caffeinated coffee consumption					
	$\leq$ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Mean	18.7	19.6	19.3	18.9	17.5	
SD	21.1	25.5	22.5	22.0	22.0	
n	12,215	6,617	17,234	12,290	2,383	50,739

(a) Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

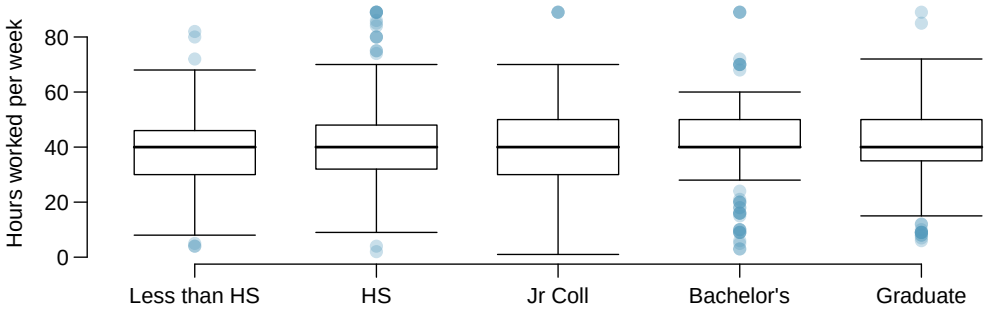
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coffee					0.0003
Residuals		25,564,819			
Total		25,575,327			

(d) What is the conclusion of the test?

[43] M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: *Archives of internal medicine* 171.17 (2011), p. 1571.

4.40 Work hours and education, Part III. In Exercises 4.16 and 4.18 you worked with data from the General Social Survey in order to compare the average number of hours worked per week by US residents with and without a college degree. However, this analysis didn't take advantage of the original data which contained more accurate information on educational attainment (less than high school, high school, junior college, Bachelor's, and graduate school). Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once instead of re-categorizing them into two groups. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

	Educational attainment					
	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172



(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

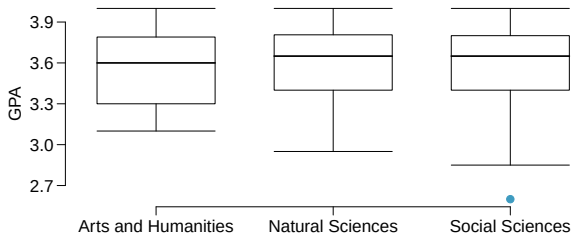
(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree			501.54		0.0682
Residuals		267,382			
Total					

(d) What is the conclusion of the test?

4.41 GPA and major. Undergraduate students taking an introductory statistics course at Duke University conducted a survey about GPA and major. The side-by-side box plots show the distribution of GPA among three groups of majors. Also provided is the ANOVA output.



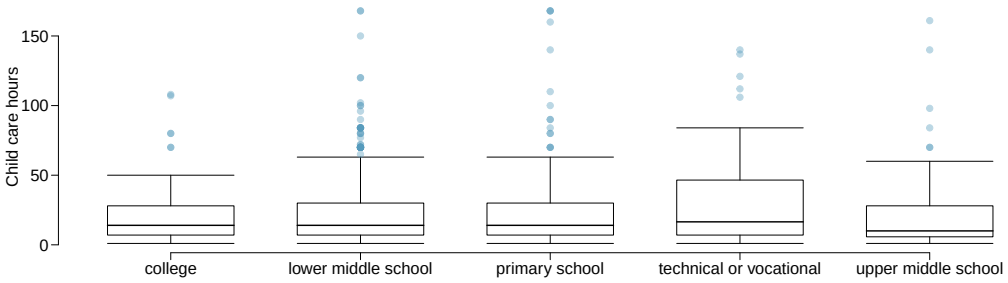
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
major	2	0.03	0.02	0.21	0.8068
Residuals	195	15.77	0.08		

(a) Write the hypotheses for testing for a difference between average GPA across majors.

(b) What is the conclusion of the hypothesis test?

(c) How many students answered these questions on the survey, i.e. what is the sample size?

4.42 Child care hours, Part II. Exercise 4.22 introduces the China Health and Nutrition Survey which, among other things, collects information on number of hours Chinese parents spend taking care of their children under age 6. The side by side box plots below show the distribution of this variable by educational attainment of the parent. Also provided below is the ANOVA output for comparing average hours across educational attainment categories.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
education	4	4142.09	1035.52	1.26	0.2846
Residuals	794	653047.83	822.48		

(a) Write the hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels.

(b) What is the conclusion of the hypothesis test?

4.43 True or false, Part II. Determine if the following statements are true or false in ANOVA, and explain your reasoning for statements you identify as false.

(a) As the number of groups increases, the modified significance level for pairwise tests increases as well.

(b) As the total sample size increases, the degrees of freedom for the residuals increases as well.

(c) The constant variance condition can be somewhat relaxed when the sample sizes are relatively consistent across groups.

(d) The independence assumption can be relaxed when the total sample size is large.

4.44 True or false, Part III. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

If the null hypothesis that the means of four groups are all the same is rejected using ANOVA at a 5% significance level, then ...

(a) we can then conclude that all the means are different from one another.

(b) the standardized variability between groups is higher than the standardized variability within groups.

(c) the pairwise analysis will identify at least one pair of means that are significantly different.

(d) the appropriate α to be used in pairwise comparisons is 0.05 / 4 = 0.0125 since there are four groups.

4.45 Prison isolation experiment, Part II. Exercise 4.35 introduced an experiment that was conducted with the goal of identifying a treatment that reduces subjects' psychopathic deviant T scores, where this score measures a person's need for control or his rebellion against control. In Exercise 4.35 you evaluated the success of each treatment individually. An alternative analysis involves comparing the success of treatments. The relevant ANOVA output is given below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	2	639.48	319.74	3.33	0.0461
Residuals	39	3740.43	95.91		

$s_{pooled} = 9.793$ on $df = 39$

(a) What are the hypotheses?

(b) What is the conclusion of the test? Use a 5% significance level.

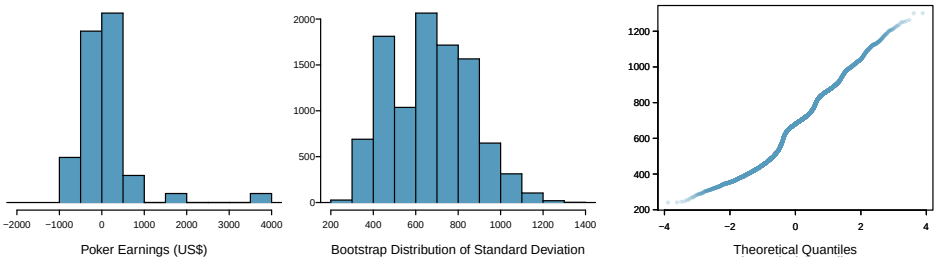
(c) If in part (b) you determined that the test is significant, conduct pairwise tests to determine which groups are different from each other. If you did not reject the null hypothesis in part (b), recheck your solution.

4.6.5 Bootstrapping to study the standard deviation

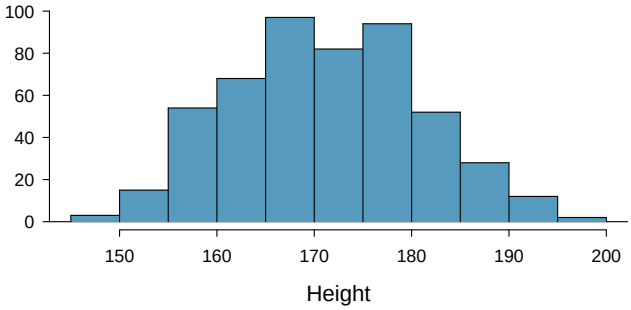
4.46 Poker winnings. An aspiring poker player recorded her winnings and losses over 50 evenings of play, summarized in the figure on the left. The daily winnings averaged \$90.08, but were very volatile with a standard deviation of \$703.68. The poker player would like to better understand how precise the standard deviation estimate is of the volatility in her long term play, so she constructed a bootstrap distribution for the standard deviation, shown on the right.

(a) Describe the distribution.

(b) Determine whether the bootstrap method is suitable for constructing a confidence interval for the standard deviation in this exercise.



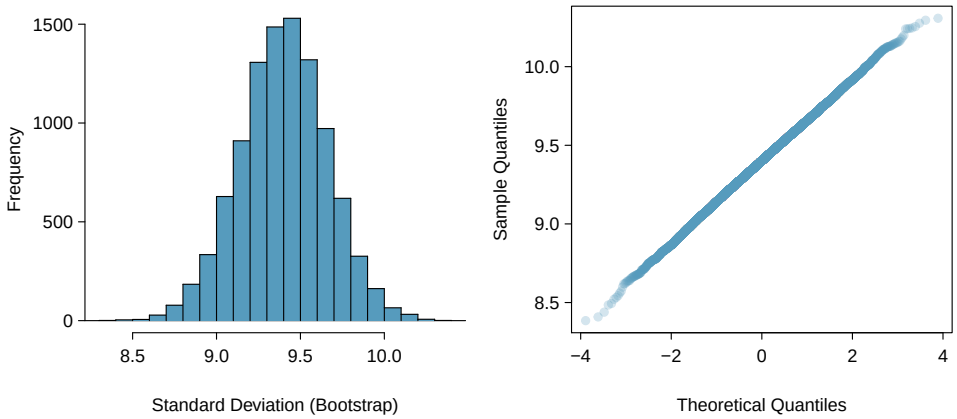
4.47 Heights of adults. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters. We would like to get 95% confidence bounds for the standard deviation of the heights in the population. For this exercise, you may assume the sample is a simple random sample from the population of interest.[44]



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.41
Q3	177.8
Max	198.1

(a) What is the point estimate for the standard deviation of the height of active individuals?

(b) The bootstrap distribution for the standard deviation is provided below. Do you think it is reasonable to construct a 95% confidence interval for the population standard deviation using the bootstrap? Explain.



(c) Below are percentiles of the bootstrap distribution. Construct a 95% confidence interval for the standard deviation.

1%	2.5%	5%	10%	50%	90%	95%	97.5%	99%
8.78	8.88	8.97	9.07	9.40	9.73	9.82	9.91	10.01

[44] G. Heinz et al. "Exploring relationships in body dimensions". In: *Journal of Statistics Education* 11.2 (2003).

Chapter 5

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 for prediction or to evaluate whether there is a linear relationship between two numerical variables.

Figure 5.1 shows two variables whose relationship can be modeled perfectly with a straight line. The equation for the line is

$$y = 5 + 57.49x$$

Imagine what a perfect linear relationship would mean: you would know the exact value of y just by knowing the value of x. This is 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 y a college may offer a prospective student. However, there would still be variability in financial support, even when comparing students whose families have similar financial backgrounds.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

$$y = \beta_0 + \beta_1 x \tag{5.1}$$

β_0, β_1
Linear model parameters

where β_0 and β_1 represent two model parameters (β is the Greek letter *beta*). These parameters are estimated using data, and we write their point estimates as b_0 and b_1. When we use x to predict y, we usually call x the explanatory or **predictor** variable, and we call y the response.

It is rare for all of the data to fall on a straight line, as seen in the three scatterplots in Figure 5.2. 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 will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less?

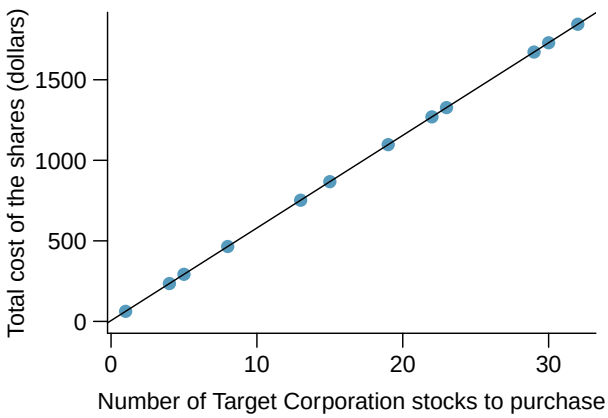


Figure 5.1: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker `TGT`, April 26th, 2012), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.

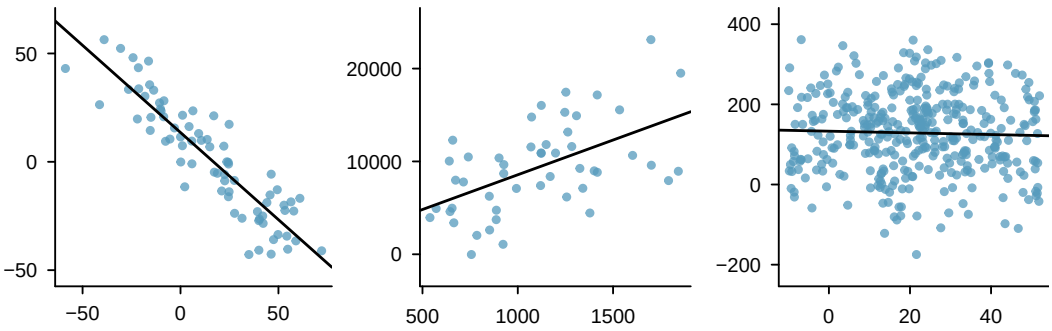


Figure 5.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

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.

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 5.3 where there is a very strong relationship between the variables even though the trend is not linear. We will discuss nonlinear trends in this chapter and the next, but the details of fitting nonlinear models are saved for a later course.

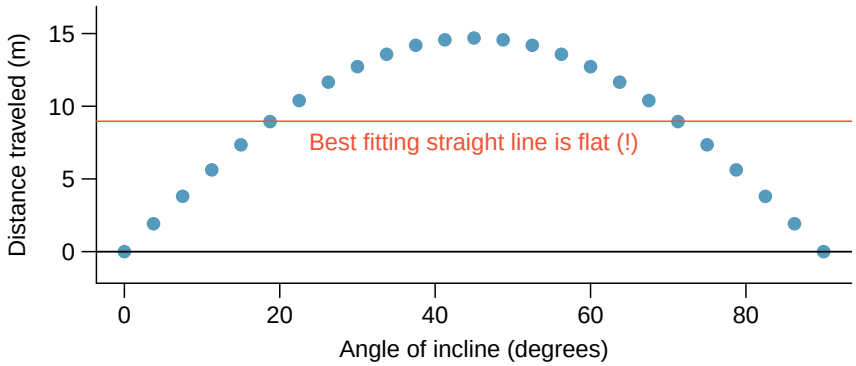


Figure 5.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

5.1 Line fitting, residuals, and correlation

It is helpful to think deeply about the line fitting process. In this section, we examine criteria for identifying a linear model and introduce a new statistic, *correlation*.

5.1.1 Beginning with straight lines

Scatterplots were introduced in Chapter 1 as a graphical technique to present two numerical variables simultaneously. Such plots permit the relationship between the variables to be examined with ease. Figure 5.4 shows a scatterplot for the head length and total length of 104 brushtail possums from Australia. Each point represents a single possum from the data.

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.

Straight lines should only be used when the data appear to have a linear relationship, such as the case shown in the left panel of Figure 5.6. The right panel of Figure 5.6 shows a case where a curved line would be more useful in understanding the relationship between the two variables.

Caution: Watch out for curved trends
We only consider models based on straight lines in this chapter. If data show a nonlinear trend, like that in the right panel of Figure 5.6, more advanced techniques should be used.

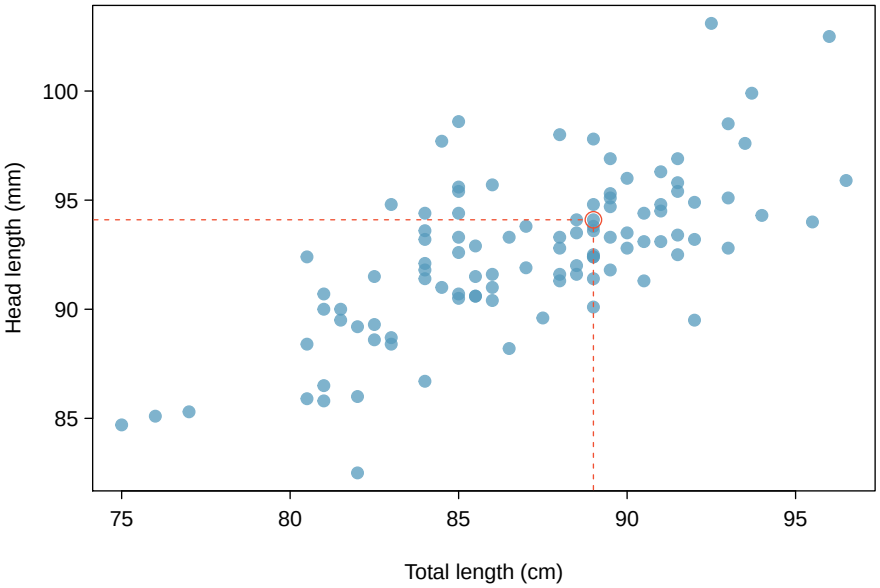


Figure 5.4: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Figure 5.5: The common brushtail possum of Australia.

Photo by wollombi on Flickr: www.flickr.com/photos/wollombi/58499575

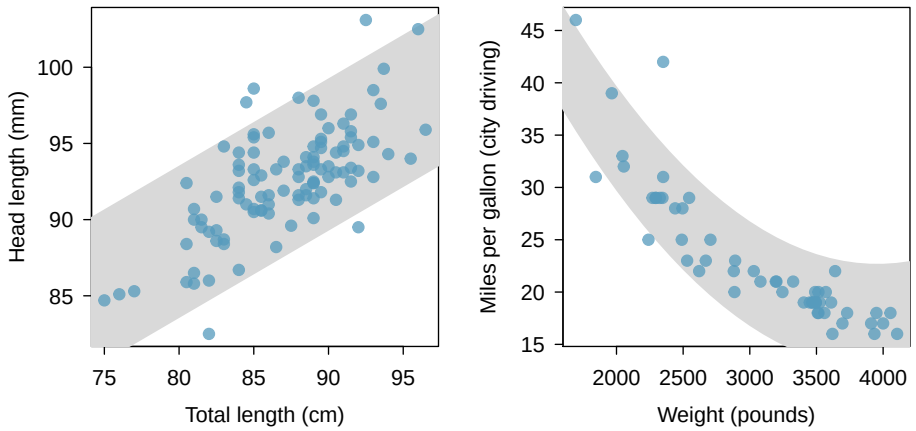


Figure 5.6: The figure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for `weight` and `mpgCity` from the `cars` data set.

5.1.2 Fitting a line by eye

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 as the predictor variable, x, to predict a possum's head length, y. We could fit the linear relationship by eye, as in Figure 5.7. The equation for this line is

$$\hat{y} = 41 + 0.59x \tag{5.2}$$

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

$$\begin{aligned} \hat{y} &= 41 + 0.59 \times 80 \\ &= 88.2 \end{aligned}$$

A "hat" on y is used to signify that this is an estimate. This estimate 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. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

5.1.3 Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Each observation will have a residual. 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.

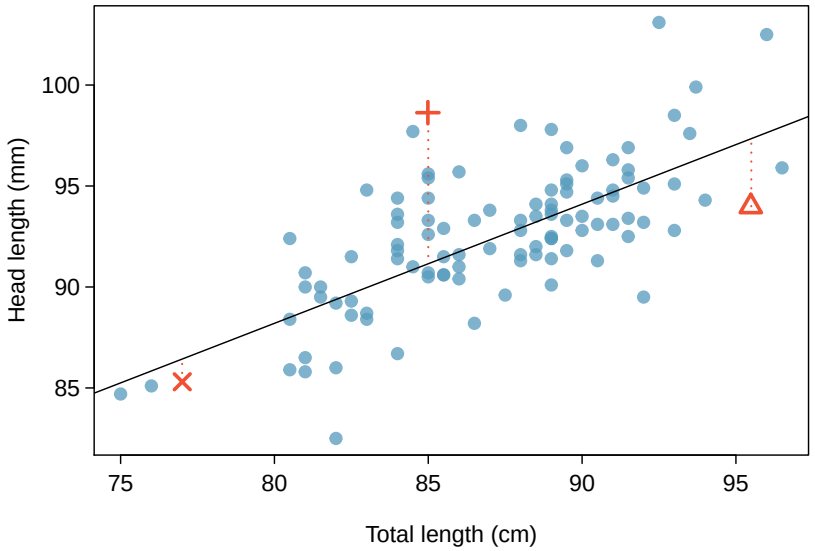


Figure 5.7: A reasonable linear model was fit to represent the relationship between head length and total length.

Three observations are noted specially in Figure 5.7. 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 of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

Example 5.3 The linear fit shown in Figure 5.7 is given as $\hat{y} = 41 + 0.59x$. Based on this line, formally compute the residual of the observation $(77.0, 85.3)$. This observation is denoted by "×" on the plot. Check it against the earlier visual estimate, -1.

We first compute the predicted value of point "×" based on the model:

$$\hat{y}_\times = 41 + 0.59x_\times = 41 + 0.59 \times 77.0 = 86.4$$

Next we compute the difference of the actual head length and the predicted head length:

$$e_\times = y_\times - \hat{y}_\times = 85.3 - 86.4 = -1.1$$

This is very close to the visual estimate of -1.

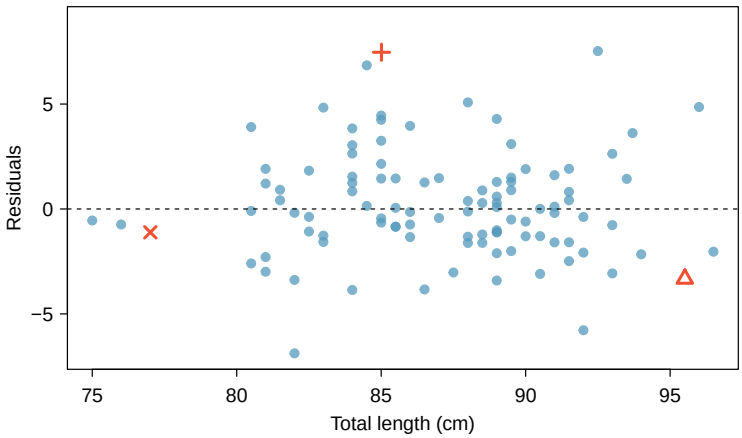


Figure 5.8: Residual plot for the model in Figure 5.7.

⊙ **Guided Practice 5.4** If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

⊙ **Guided Practice 5.5** Compute the residuals for the observations $(85.0, 98.6)$ ("+" in the figure) and $(95.5, 94.0)$ ("△") using the linear relationship $\hat{y} = 41 + 0.59x$. [2]

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a **residual plot** such as the one shown in Figure 5.8 for the regression line in Figure 5.7. The residuals are plotted at their original horizontal locations but with the vertical coordinate as the residual. 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.

● **Example 5.6** One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 5.9 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.

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.

[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:

$$\hat{y}_+ = 41 + 0.59x_+ = 41 + 0.59 \times 85.0 = 91.15$$

Then the residual is given by

$$e_+ = y_+ - \hat{y}_+ = 98.6 - 91.15 = 7.45$$

This was close to the earlier estimate of 7.

(△) $\hat{y}_\triangle = 41 + 0.59x_\triangle = 97.3$. $e_\triangle = y_\triangle - \hat{y}_\triangle = -3.3$, close to the estimate of -4.

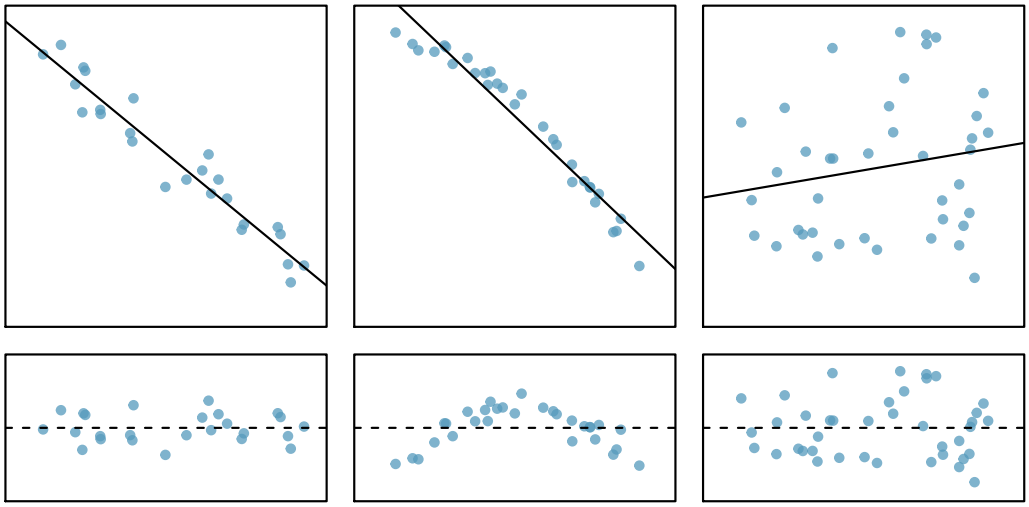


Figure 5.9: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

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 point estimate of the slope parameter, labeled b_1, is not zero, but we might wonder if this could just be due to chance. We will address this sort of scenario in Section 5.4.

5.1.4 Describing linear relationships with correlation

Correlation: strength of a linear relationship

R
correlation

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

We can compute the correlation using a formula, just as we did with the sample mean and standard deviation. However, this formula is rather complex,[3] so we generally perform the calculations on a computer or calculator. Figure 5.10 shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the 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.

The correlation is intended to quantify the strength of a linear trend. Nonlinear trends,

[3]Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) using the formula

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

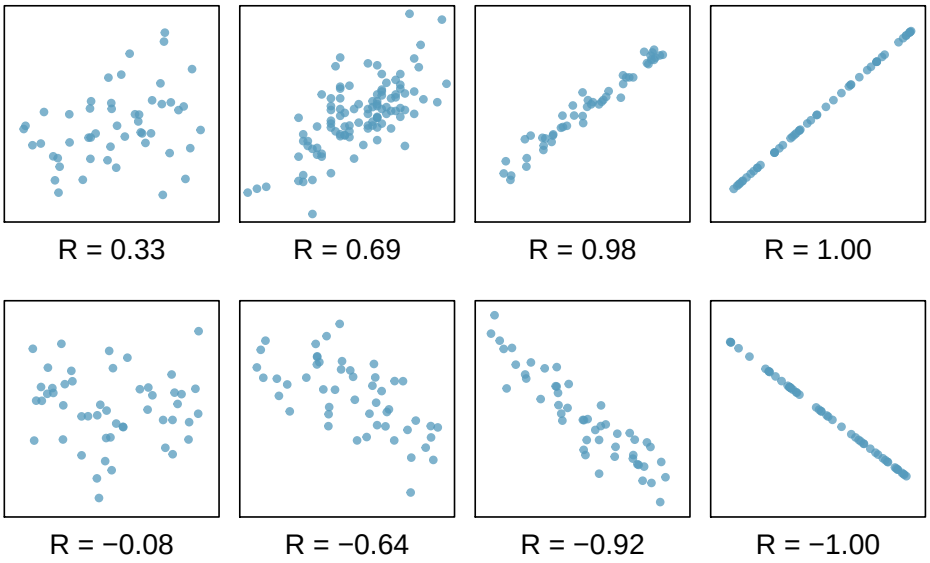


Figure 5.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 5.11.

⊙ **Guided Practice 5.7** It appears no straight line would fit any of the datasets represented in Figure 5.11. Instead, try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

5.2 Fitting a line by least squares regression

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 2011 freshman class of Elmhurst College in Illinois.[5] 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 5.12 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.

⊙ **Guided Practice 5.8** Is the correlation positive or negative in Figure 5.12?[6]

[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.

[5] 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

[6] Larger family incomes are associated with lower amounts of aid, so the correlation will be negative. Using a computer, the correlation can be computed: -0.499.

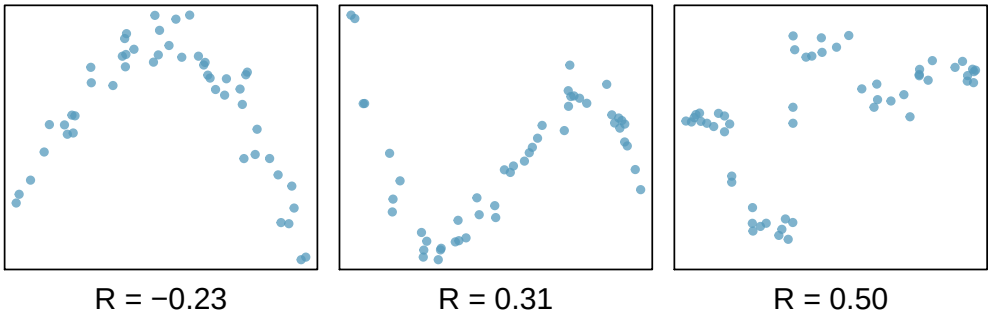


Figure 5.11: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

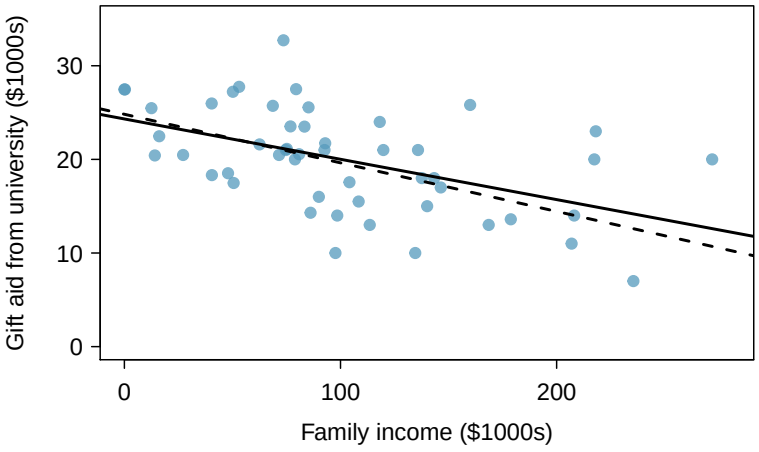


Figure 5.12: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

5.2.1 An objective measure for finding the best line

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:

$$|e_1| + |e_2| + \cdots + |e_n| \tag{5.9}$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 5.12 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:

$$e_1^2 + e_2^2 + \cdots + e_n^2 \tag{5.10}$$

The line that minimizes this **least squares criterion** is represented as the solid line in Figure 5.12. This is commonly called the **least squares line**. The following are three possible reasons to choose Criterion (5.10) over Criterion (5.9):

1. It is the most commonly used method.
2. Computing the line based on Criterion (5.10) is much easier by hand and in most statistical software.
3. 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.

The first two reasons are largely for tradition and convenience; the last reason explains why Criterion (5.10) is typically most helpful.[7]

5.2.2 Finding the least squares line

For the Elmhurst data, we could write the equation of the least squares regression line as

$$\widehat{aid} = \beta_0 + \beta_1 \times family\_income$$

Here the equation is set up to predict gift aid based on a student's family income, which would be useful to students considering Elmhurst. These two values, β_0 and β_1, are the *parameters* of the regression line.

As in Chapters 4-6, the parameters are estimated using observed data. In practice, this estimation is done using a computer in the same way that other estimates, like a sample mean, can be estimated using a computer or calculator. However, we can also find the parameter estimates by applying two properties of the least squares line:

- The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} R \tag{5.11}$$

  where R is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable and response, respectively.
- If $\bar{x}$ is the mean of the horizontal variable (from the data) and $\bar{y}$ is the mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line.

We use b_0 and b_1 to represent the point estimates of the parameters β_0 and β_1.

b_0, b_1
Sample estimates of β_0, β_1

[7] There are applications where Criterion (5.9) may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

⊙ **Guided Practice 5.12** Table 5.13 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 5.12 on page 228 to verify it falls on the least squares line (the solid line).[8]

	family income, in \$1000s ("x")	gift aid, in \$1000s ("y")
mean	$\bar{x} = 101.8$	$\bar{y} = 19.94$
sd	$s_x = 63.2$	$s_y = 5.46$
		$R = -0.499$

Table 5.13: Summary statistics for family income and gift aid.

⊙ **Guided Practice 5.13** Using the summary statistics in Table 5.13, compute the slope for the regression line of gift aid against family income.[9]

You might recall the **point-slope** form of a line from math class (another common form is *slope-intercept*). Given the slope of a line and a point on the line, (x_0, y_0), the equation for the line can be written as

$$y - y_0 = slope \times (x - x_0) \tag{5.14}$$

A common exercise to become more familiar with foundations of least squares regression is to use basic summary statistics and point-slope form to produce the least squares line.

TIP: Identifying the least squares line from summary statistics
To identify the least squares line from summary statistics:

- Estimate the slope parameter, β_1, by calculating b_1 using Equation (5.11).
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ along with the slope b_1 in the point-slope equation:

$$y - \bar{y} = b_1(x - \bar{x})$$

- Simplify the equation.

[8] 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.

[9] Apply Equation (5.11) with the summary statistics from Table 5.13 to compute the slope:

$$b_1 = \frac{s_y}{s_x} R = \frac{5.46}{63.2}(-0.499) = -0.0431$$

Example 5.15 Using the point $(101.8, 19.94)$ from the sample means and the slope estimate $b_1 = -0.0431$ from Guided Practice 5.13, find the least-squares line for predicting aid based on family income.

Apply the point-slope equation using $(101.8, 19.94)$ and the slope $b_1 = -0.0431$:

$$y - y_0 = b_1(x - x_0)$$
$$y - 19.94 = -0.0431(x - 101.8)$$

Expanding the right side and then adding 19.94 to each side, the equation simplifies:

$$\widehat{aid} = 24.3 - 0.0431 \times family\_income$$

Here we have replaced y with $\widehat{aid}$ and x with $family\_income$ to put the equation in context.

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 Table 5.14 for the Elmhurst data. The first column of numbers provides estimates for b_0 and b_1, respectively. Compare these to the result from Example 5.15.

	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

Table 5.14: Summary of least squares fit for the Elmhurst data. Compare the parameter estimates in the first column to the results of Example 5.15.

Example 5.16 Examine the second, third, and fourth columns in Table 5.14. Can you guess what they represent?

We'll describe the meaning of the columns using the second row, which corresponds to β_1. The first column provides the point estimate for β_1, as we calculated in an earlier example: -0.0431. The second column is a standard error for this point estimate: 0.0108. The third column is a t test statistic for the null hypothesis that $\beta_1 = 0$: $T = -3.98$. The last column is the p-value for the t test statistic for the null hypothesis $\beta_1 = 0$ and a two-sided alternative hypothesis: 0.0002. We will get into more of these details in Section 5.4.

Example 5.17 Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have estimated to calculate her financial aid from the university?

She may use it as an estimate, though some qualifiers on this approach are important. First, the data all come from one freshman class, and the way aid is determined by the university may change from year to year. Second, the equation will provide an imperfect estimate. While the linear equation is good at capturing the trend in the data, no individual student's aid will be perfectly predicted.

5.2.3 Interpreting regression line parameter estimates

Interpreting parameters in a regression model is often one of the most important steps in the analysis.

Example 5.18 The slope and intercept estimates for the Elmhurst data are -0.0431 and 24.3. What do these numbers really mean?

Interpreting the slope parameter is helpful in almost any application. For each additional \$1,000 of family income, we would expect a student to receive a net difference of \$1,000 $\times$ (-0.0431) = $-$\$43.10 in aid on average, i.e. \$43.10 *less*. Note that a higher family income corresponds to less aid because the coefficient of family income is negative in the model. We must be cautious in this interpretation: while there is a real association, we cannot interpret a causal connection between the variables because these data are observational. That is, increasing a student's family income may not cause the student's aid to drop. (It would be reasonable to contact the college and ask if the relationship is causal, i.e. if Elmhurst College's aid decisions are partially based on students' family income.)

The estimated intercept $b_0 = 24.3$ (in \$1000s) describes the average aid if a student's family had 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.

Interpreting parameters estimated by least squares
The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if $x = 0$ *and* the linear model is valid all the way to $x = 0$, which in many applications is not the case.

5.2.4 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 6^{th} 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 [10]

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] http://www.colbertnation.com/the-colbert-report-videos/269929/

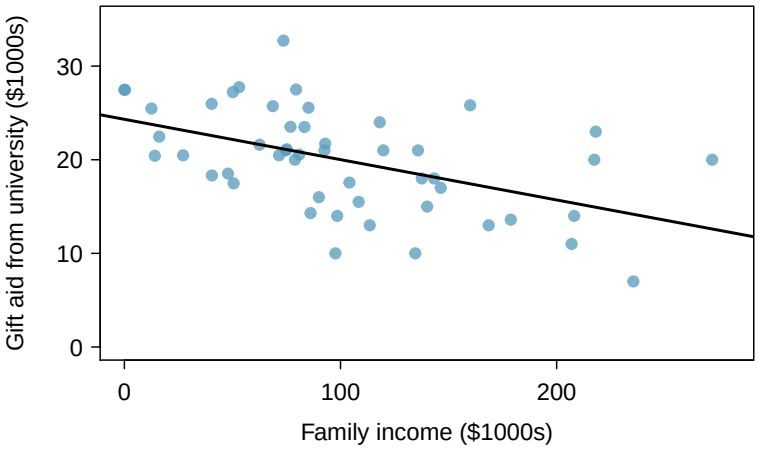


Figure 5.15: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

Example 5.19 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 $family\_income = 1000$:

$$24.3 - 0.0431 \times family\_income = 24.3 - 0.0431 \times 1000 = -18.8$$

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.

Applying a model estimate to values outside of the realm 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 explored.

5.2.5 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 strength of a linear fit using R^2, called **R-squared**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line. For example, consider the Elmhurst data, shown in Figure 5.15. 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: $s^2_{RES} = 22.4$. In short, there was a reduction of

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29.8 - 22.4}{29.8} = \frac{7.5}{29.8} = 0.25$$

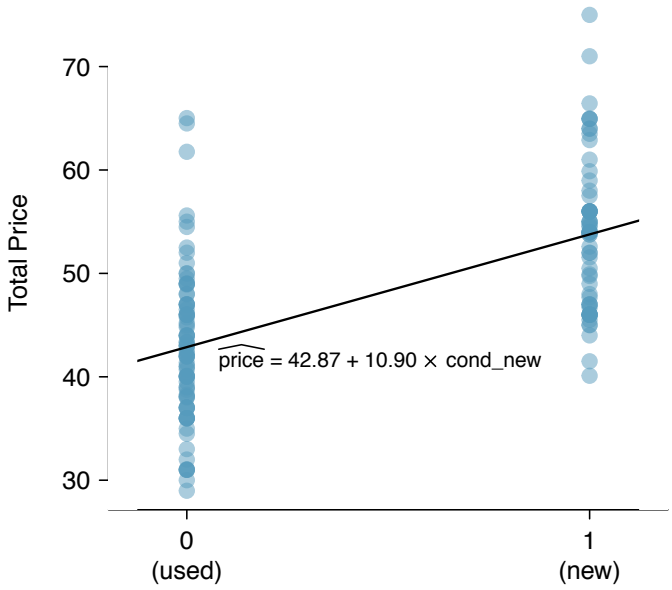


Figure 5.16: Total auction prices for the video game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

or about 25% in the data's variation by using information about family income for predicting aid using a linear model. This corresponds exactly to the R-squared value:

$$R = -0.499 \qquad\qquad R^2 = 0.25$$

⊙ **Guided Practice 5.20** If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?[11]

5.2.6 Categorical predictors with two levels

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.[12] 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 5.16.

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

$$\widehat{price} = \beta_0 + \beta_1 \times \texttt{cond\_new}$$

[11] About $R^2 = (-0.97)^2 = 0.94$ or 94% of the variation is explained by the linear model.
[12] These data were collected in Fall 2009 and may be found at openintro.org.

	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

Table 5.17: Least squares regression summary for the final auction price against the condition of the game.

The fitted model is summarized in Table 5.17, and the model with its parameter estimates is given as

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

Example 5.21 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.

TIP: 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.

We'll elaborate further on this Ebay auction data in Chapter 6, where we examine the influence of many predictor variables simultaneously using multiple regression. In multiple regression, we will consider the association of auction price with regard to each variable while controlling for the influence of other variables. This is especially important since some of the predictors are associated. For example, auctions with games in new condition also often came with more accessories.

5.3 Types of outliers in linear regression

In this section, we identify criteria for determining which outliers are important and influential.

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 5.22 There are six plots shown in Figure 5.18 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 square 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 5.18. 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 5.22 – 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.

Caution: 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.

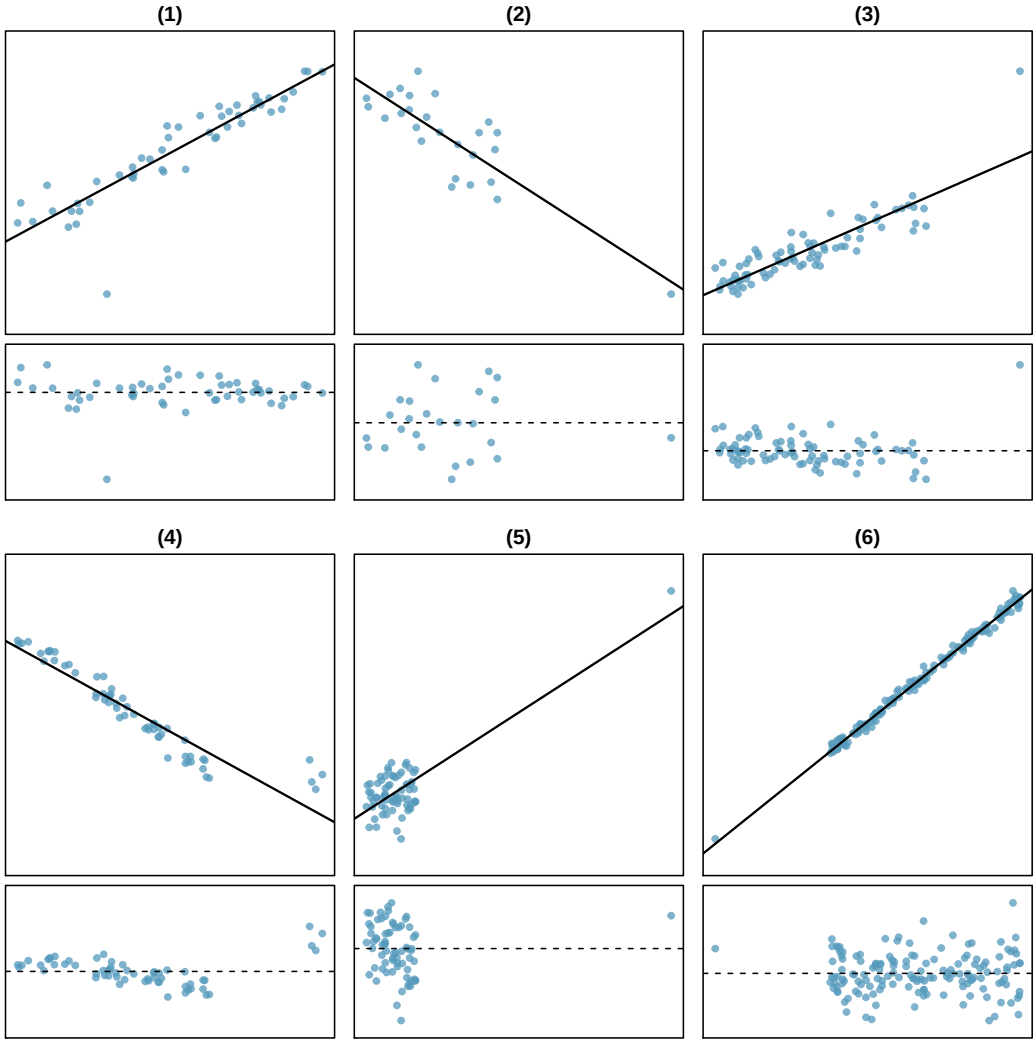


Figure 5.18: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

Caution: Outliers for a categorical predictor with two levels
Be cautious about using a categorical predictor when one of the levels has very few observations. When this happens, those few observations become influential points.

5.4 Inference for linear regression

In this section we discuss uncertainty in the estimates of the slope and y-intercept for a regression line. Just as we identified standard errors for point estimates in previous chapters, we first discuss standard errors for these new estimates. However, in the case of regression, we will identify standard errors using statistical software.

5.4.1 Conditions for the least squares line

When performing inference on a least squares line, we generally require the following:

Linearity. The data should show a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 5.19), an advanced regression method from another book or later course should be applied.

Nearly normal residuals. Generally the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points, which we will discuss in greater depth in Section 5.3. An example of non-normal residuals is shown in the second panel of Figure 5.19.

Constant variability. The variability of points around the least squares line remains roughly constant. An example of non-constant variability is shown in the third panel of Figure 5.19.

Independent observations. Be cautious about applying regression to data collected sequentially in what is called a **time series**. Such data may have an underlying structure that should be considered in a model and analysis. An example of a time series where independence is violated is shown in the fourth panel of Figure 5.19.

For additional information on checking regression conditions, see Section 6.3.

Example 5.23 Should we have concerns about applying inference to the Elmhurst data in Figure 5.20?

The trend appears to be linear, the data fall around the line with no obvious outliers, the variance is roughly constant. These are also not time series observations. It would be reasonable to analyze the model using inference.

5.4.2 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 2010, with the exception

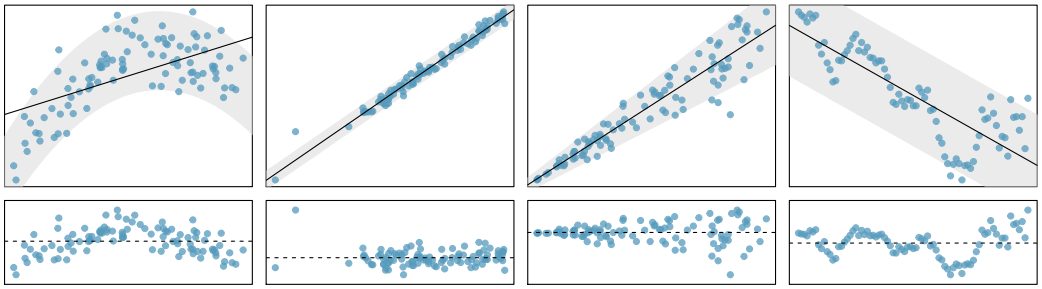


Figure 5.19: Four examples showing when the methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not fit the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points 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.

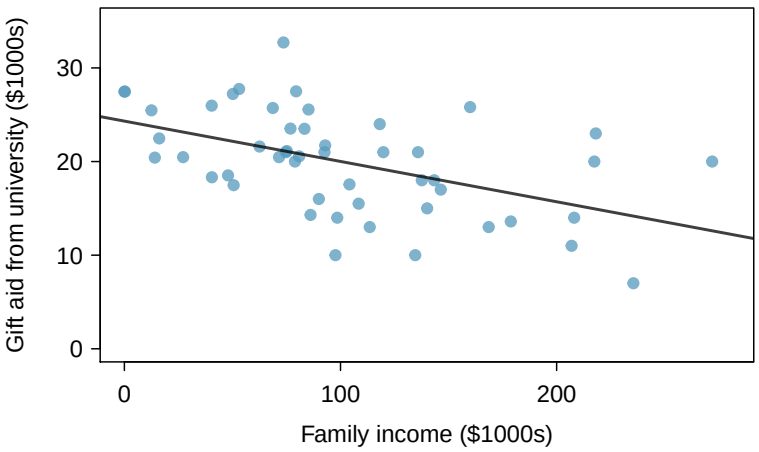


Figure 5.20: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

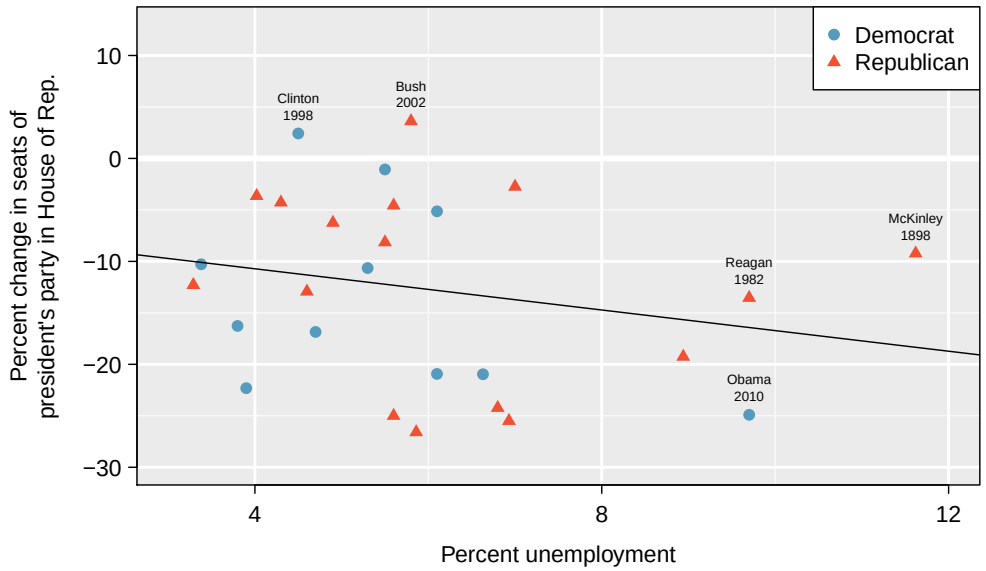


Figure 5.21: The percent change in House seats for the President's party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data.

of those elections during the Great Depression. Figure 5.21 shows these data and the least-squares regression line:

$$\begin{aligned}&\%\text{ change in House seats for President's party}\\&\quad = -6.71 - 1.00 \times (\text{unemployment rate})\end{aligned}$$

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 Democrats in 2010) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or in the normality of residuals (though we don't examine a normal probability plot here). 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.

⊙ **Guided Practice 5.24** 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?[13]

There is a negative slope in the line shown in Figure 5.21. 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

[13] 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.

investigation into a two-sided statistical hypothesis test. We use a two-sided test since a statistically significant result in either direction would be interesting.

H_0: $\beta_1 = 0$. The true linear model has slope zero.

H_A: $\beta_1 \neq 0$. The true linear model has a slope different than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives, or vice-versa.

We would reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter 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.

5.4.3 Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b_1. We will generally label the test statistic using a T, since it follows the t distribution.

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. Table 5.22 shows software output for the least squares regression line in Figure 5.21. 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)	-6.7142	5.4567	-1.23	0.2300
unemp	-1.0010	0.8717	-1.15	0.2617
				$df = 25$

Table 5.22: Output from statistical software for the regression line modeling the midterm election gains and losses for the President's party as a response to unemployment.

Example 5.25 What do the first and second columns of Table 5.22 represent?

The entries in the first column represent the least squares estimates, b_0 and b_1, and the values in the second column correspond to the standard errors of each estimate.

We previously used a t test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can compute the test statistic using the T (or Z) score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-1.0010 - 0}{0.8717} = -1.15$$

We can look for the two-tailed p-value – shown in Figure 5.23 – using the probability table for the t distribution in Appendix C.2 on page 342.

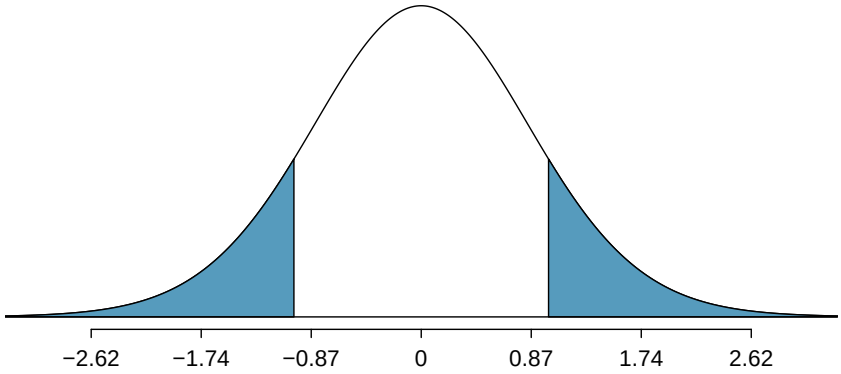


Figure 5.23: The distribution shown here is the sampling distribution for b_1, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President's party during a midterm election.

Example 5.26 Table 5.22 offers the degrees of freedom for the test statistic T: $df = 25$. Identify the p-value for the hypothesis test.

Looking in the 25 degrees of freedom row in Appendix C.2, we see that the absolute value of the test statistic is smaller than any value listed, which means the tail area and therefore also the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we fail to reject the null hypothesis. That is, the data do not provide convincing evidence that unemployment is a good predictor of how well a president's party will do in the midterm elections for the House of Representatives.

We could have identified the t test statistic from the software output in Table 5.22, shown in the second row (unemp) and third column (t value). The entry in the second row and last column in Table 5.22 represents the p-value for the two-sided hypothesis test where the null value is zero.

Inference for regression

We usually rely on statistical software to identify point estimates and standard errors for parameters of a regression line. After verifying conditions hold for fitting a line, we can use the methods learned in Section 4.1 for the t distribution to create confidence intervals for regression parameters or to evaluate hypothesis tests.

Caution: 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 a hypothesis test should be one-sided or a comparison is being made to a value other than zero, be cautious about using the software output to obtain the p-value.

Example 5.27 Examine Figure 5.15 on page 233, which relates the Elmhurst College aid and student family income. How sure are you that the slope is statistically significantly different from zero? That is, do you think a formal hypothesis test would reject the claim that the true slope of the line should be zero?

While the relationship between the variables is not perfect, there is an evident decreasing trend in the data. This suggests the hypothesis test will reject the null claim that the slope is zero.

Guided Practice 5.28 Table 5.24 shows statistical software output from fitting the least squares regression line shown in Figure 5.15. Use this output to formally evaluate the following hypotheses. H_0: The true coefficient for family income is zero. H_A: The true coefficient for family income is not zero.[14]

	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
				$df = 48$

Table 5.24: Summary of least squares fit for the Elmhurst College data.

TIP: Always check assumptions
If conditions for fitting the regression line do not hold, then the methods presented here should not be applied. The standard error or distribution assumption of the point estimate – assumed to be normal when applying the t test statistic – may not be valid.

5.4.4 An alternative test statistic

We considered the t test statistic as a way to evaluate the strength of evidence for a hypothesis test in Section 5.4.3. However, we could focus on R^2. Recall that R^2 described the proportion of variability in the response variable (y) explained by the explanatory variable (x). If this proportion is large, then this suggests a linear relationship exists between the variables. If this proportion is small, then the evidence provided by the data may not be convincing.

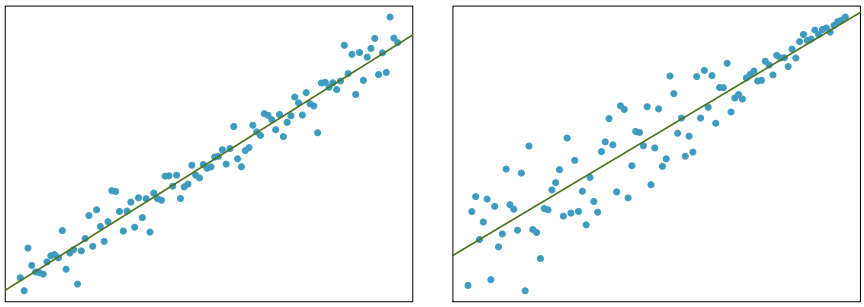
This concept – considering the amount of variability in the response variable explained by the explanatory variable – is a key component in some statistical techniques. The *analysis of variance (ANOVA)* technique introduced in Section 4.4 uses this general principle. The method states that if enough variability is explained away by the categories, then we would conclude the mean varied between the categories. On the other hand, we might not be convinced if only a little variability is explained. ANOVA can be further employed in advanced regression modeling to evaluate the inclusion of explanatory variables, though we leave these details to a later course.

[14] We look in the second row corresponding to the family income variable. We see the point estimate of the slope of the line is -0.0431, the standard error of this estimate is 0.0108, and the t test statistic is -3.98. The p-value corresponds exactly to the two-sided test we are interested in: 0.0002. The p-value is so small that we reject the null hypothesis and conclude that family income and financial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated and the true slope parameter is indeed less than 0, just as we believed in Example 5.27.

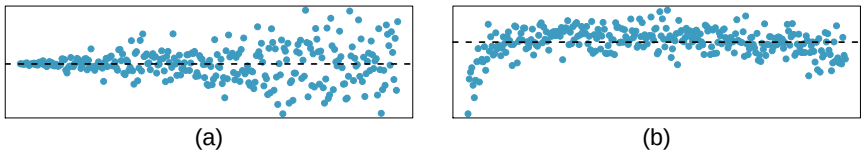
5.5 Exercises

5.5.1 Line fitting, residuals, and correlation

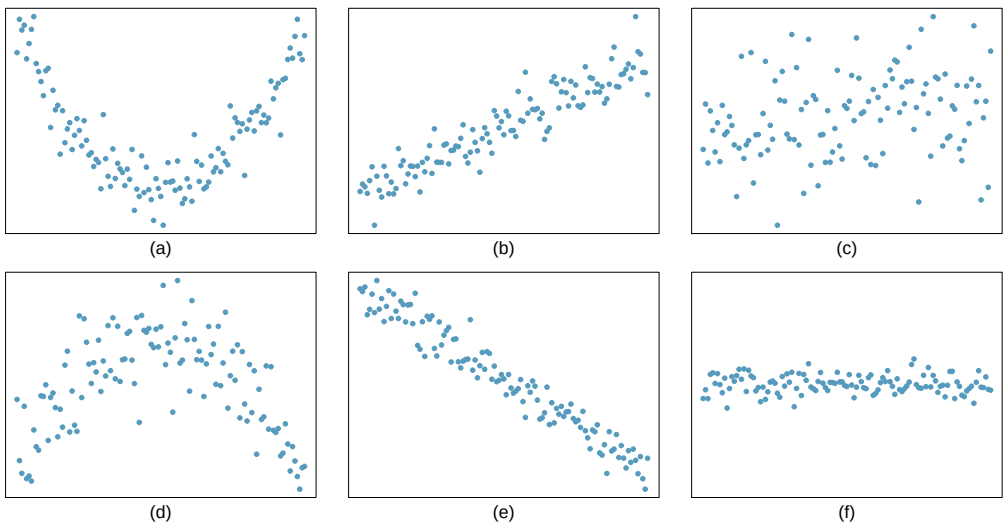
5.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.



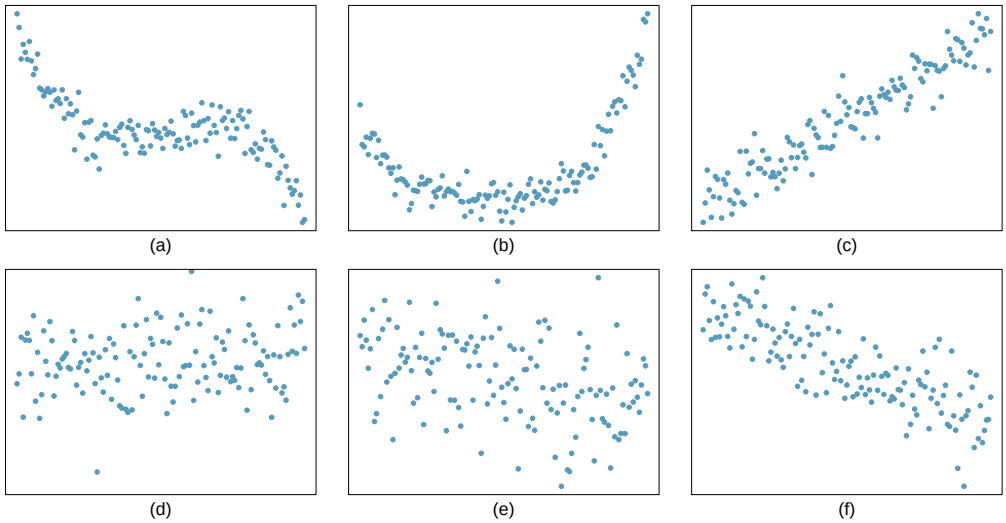
5.2 Trends in the residuals. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.



5.3 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.



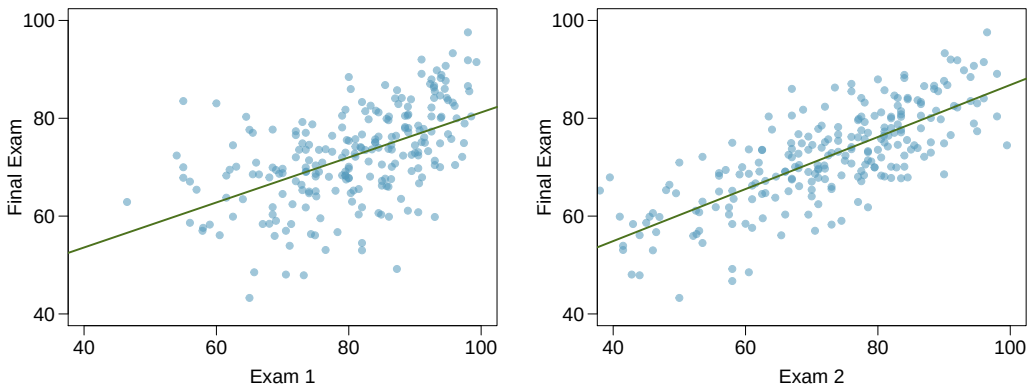
5.4 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



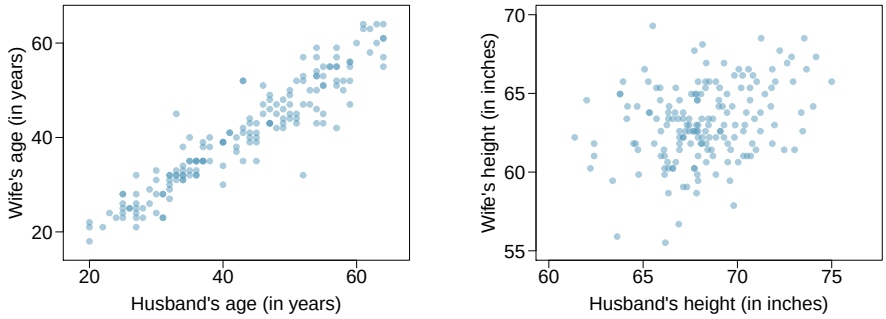
5.5 Exams and grades. The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

(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?



5.6 Husbands and wives, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives.[15] The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.



(a) Describe the relationship between husbands' and wives' ages.

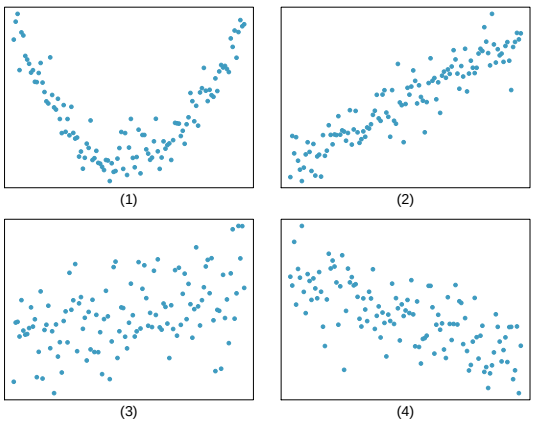
(b) Describe the relationship between husbands' and wives' heights.

(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 husbands' and wives' heights?

5.7 Match the correlation, Part I. Match the calculated correlations to the corresponding scatterplot.

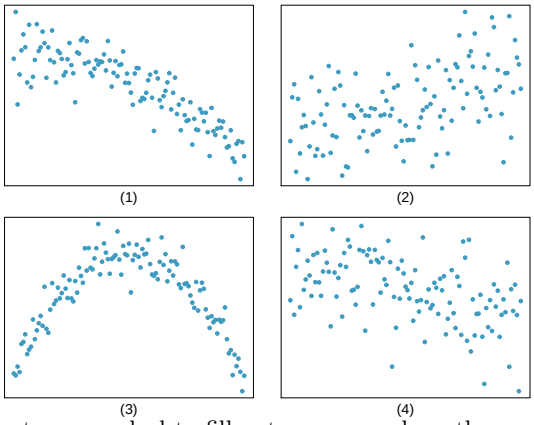
(a) $R = -0.7$
(b) $R = 0.45$
(c) $R = 0.06$
(d) $R = 0.92$



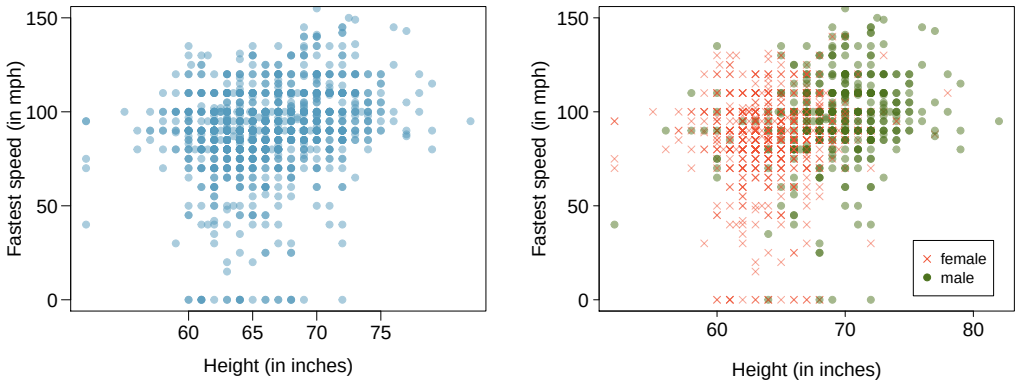
[15] D.J. Hand. *A handbook of small data sets.* Chapman & Hall/CRC, 1994.

5.8 Match the correlation, Part II. Match the calculated correlations to the corresponding scatterplot.

(a) $R = 0.49$
(b) $R = -0.48$
(c) $R = -0.03$
(d) $R = -0.85$

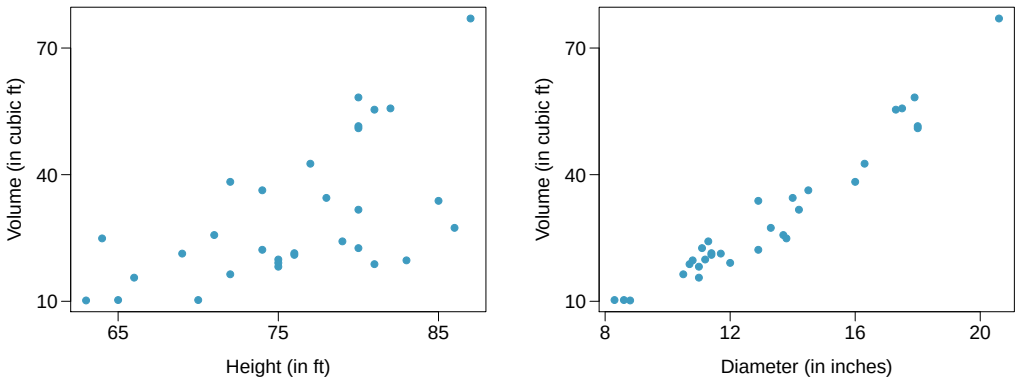


5.9 Speed and height. 1,302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown 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?

5.10 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.[16]



(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

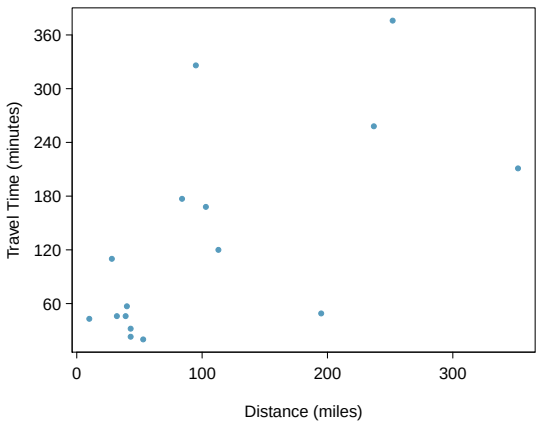
(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

5.11 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)?



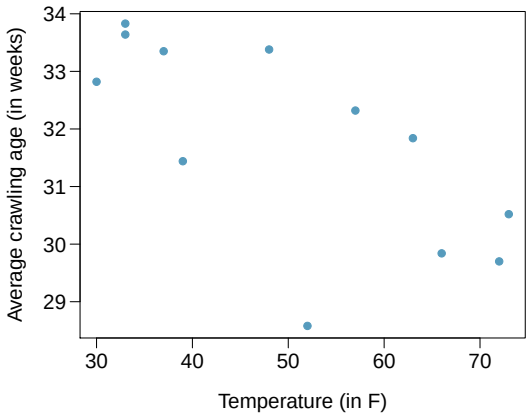
[16] Source: R Dataset, http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html.

5.12 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.[17] 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?

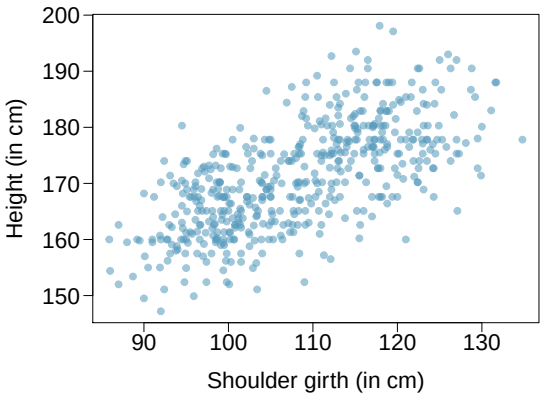
(c) The correlation between temperature in °F and age in weeks was $R = -0.70$. If we converted the temperature to °C and age to months, what would the correlation be?



5.13 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.[18] 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?



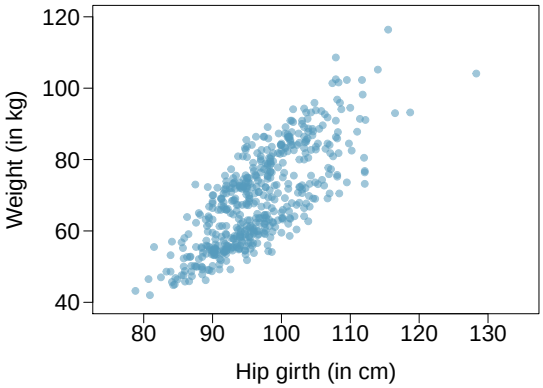
[17] J.B. Benson. "Season of birth and onset of locomotion: Theoretical and methodological implications". In: *Infant behavior and development* 16.1 (1993), pp. 69–81. ISSN: 0163-6383.

[18] G. Heinz et al. "Exploring relationships in body dimensions". In: *Journal of Statistics Education* 11.2 (2003).

5.14 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 5.13.

(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?



5.15 Correlation, Part I. What would be the correlation between the ages of husbands and wives if men always married woman who were

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

5.16 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?

5.5.2 Fitting a line by least squares regression

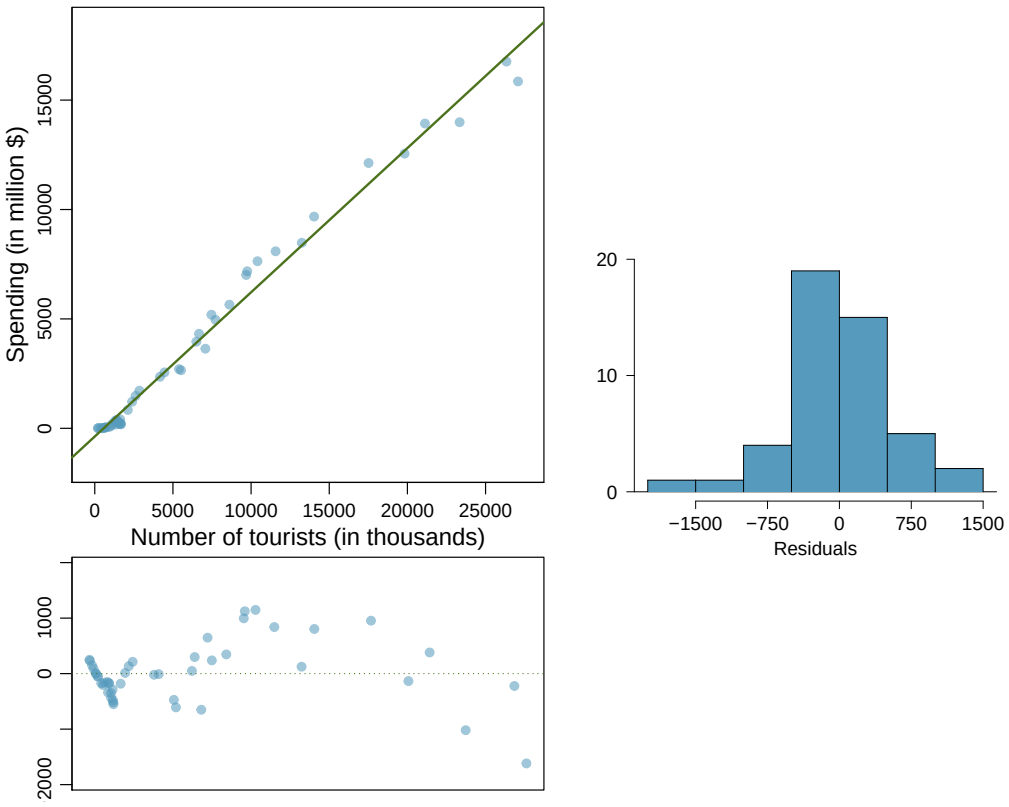
5.17 Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.[19] The scatterplot below shows the relationship between these two variables along with the least squares fit.

(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.

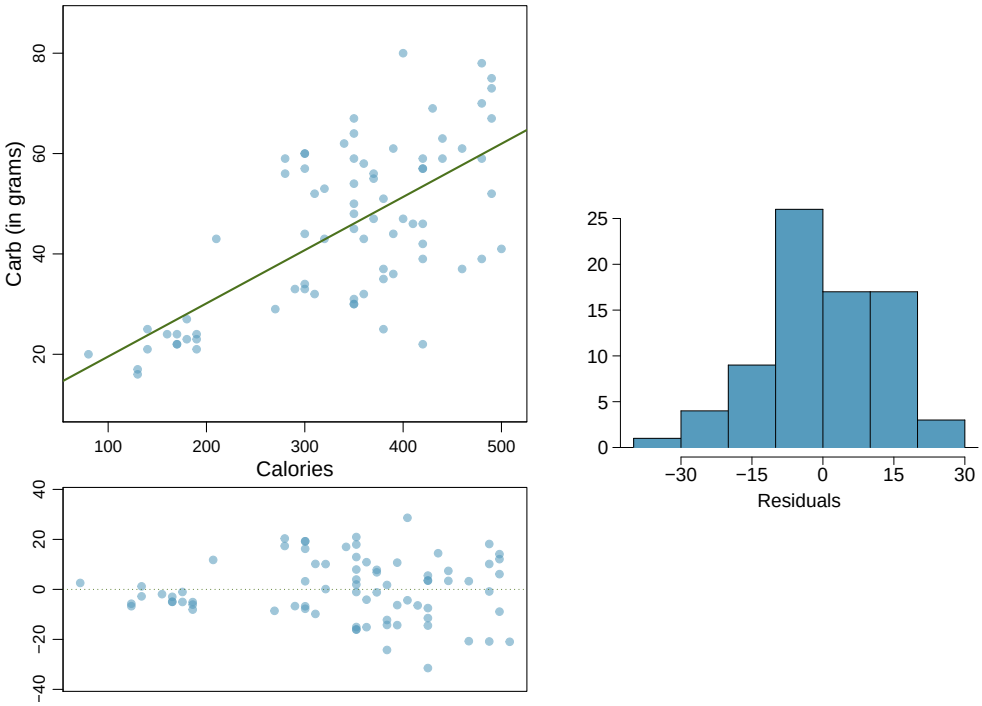
5.18 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.[20] Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount 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?



5.19 The Coast Starlight, Part II. Exercise 5.11 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 107 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.

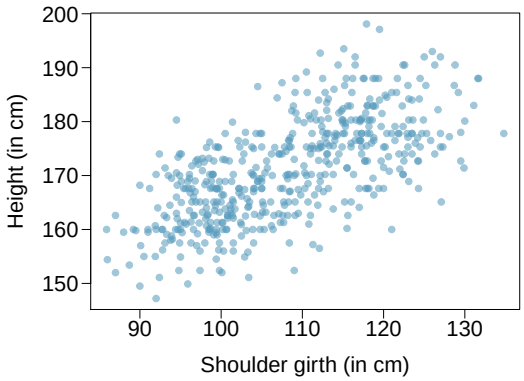
(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(e) It actually takes the 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?

[20] Source: Starbucks.com, collected on March 10, 2011, http://www.starbucks.com/menu/nutrition.

5.20 Body measurements, Part III. Exercise 5.13 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 108.20 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?

5.21 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%.

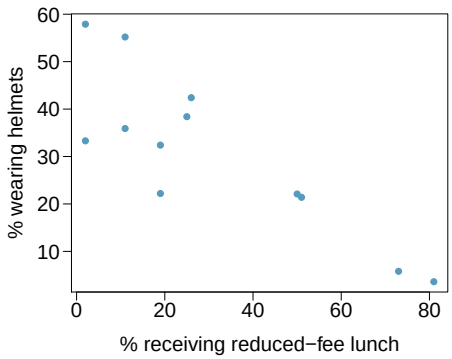
(a) If the R^2 for the least-squares regression line for these data is 72%, what is the correlation between `lunch` and `helmet`?

(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.

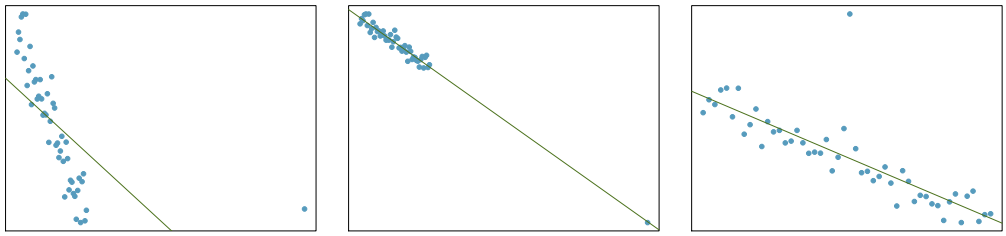
(d) Interpret the slope of the least-squares regression line in the context of the application.

(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.

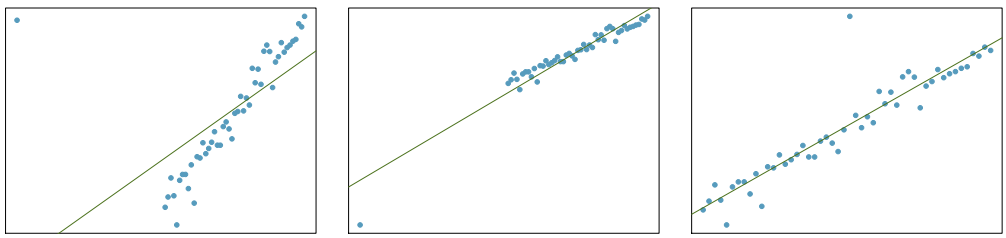


5.5.3 Types of outliers in linear regression

5.22 Outliers, Part I. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.



5.23 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.

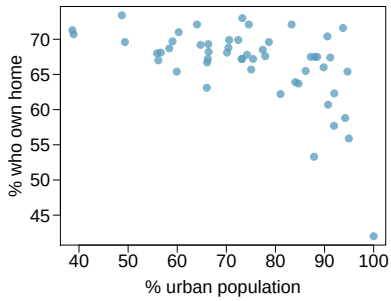


5.24 Crawling babies, Part II. Exercise 5.12 introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying 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?

5.25 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 in 2010.[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 in 2010.

(b) The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of outlier is this observation?

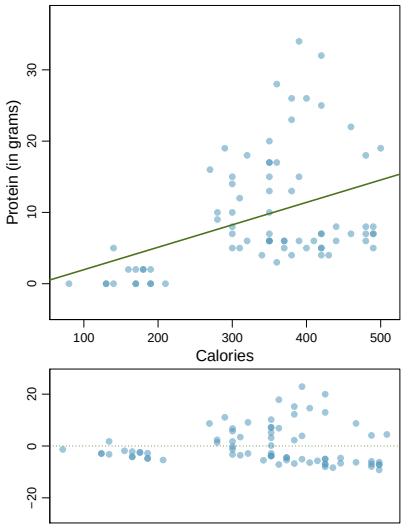


5.5.4 Inference for linear regression

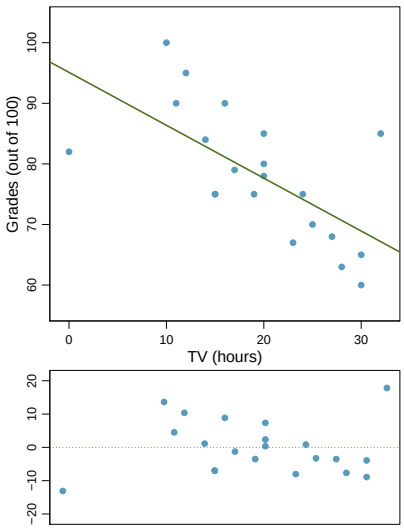
Visually check the conditions for fitting a least squares regression line, but you do not need to report these conditions in your solutions unless it is requested.

[21] United States Census Bureau, 2010 Census Urban and Rural Classification and Urban Area Criteria and Housing Characteristics: 2010.

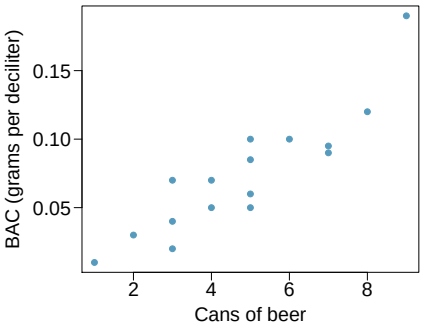
5.26 Nutrition at Starbucks, Part II. Exercise 5.18 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.



5.27 Grades and TV. Data were collected on the number of hours per week students watch TV and the grade they earned in a biology class on a 100 point scale. Based on the scatterplot and the residual plot provided, describe the relationship between the two variables, and determine if a simple linear model is appropriate to predict a student's grade from the number of hours per week the student watches TV.



5.28 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 evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.[22] 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.

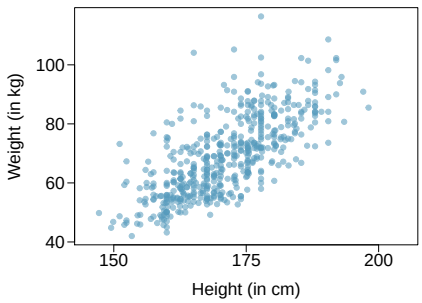
(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(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?

5.29 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.

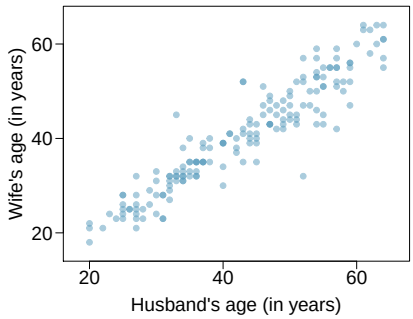
(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(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.

[22] J. Malkevitch and L.M. Lesser. *For All Practical Purposes: Mathematical Literacy in Today's World.* WH Freeman & Co, 2008.

5.30 Husbands and wives, Part II. Exercise 5.6 presents a scatterplot displaying the relationship between husbands' and wives' ages in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Given below is summary output of the least squares fit for predicting wife's age from husband's age.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5740	1.1501	1.37	0.1730
age_husband	0.9112	0.0259	35.25	0.0000

$df = 168$

(a) We might wonder, is the age difference between husbands and wives consistent across ages? If this were the case, then the slope parameter would be $\beta_1 = 1$. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages differs for different ages.

(b) Write the equation of the regression line for predicting wife's age from husband's age.

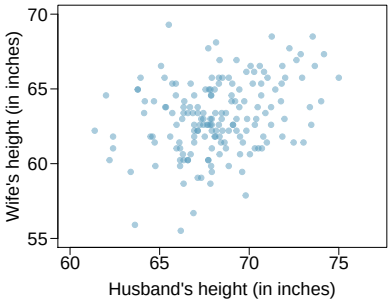
(c) Interpret the slope and intercept in context.

(d) Given that $R^2 = 0.88$, what is the correlation of ages in this data set?

(e) You meet a married man from Britain who is 55 years old. What would you predict his wife's age to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife's age? Explain.

5.31 Husbands and wives, Part III. The scatterplot below summarizes husbands' and wives' heights in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Summary output of the least squares fit for predicting wife's height from husband'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_husband	0.2863	0.0686	4.17	0.0000

(a) Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test.

(b) Write the equation of the regression line for predicting wife's height from husband'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?

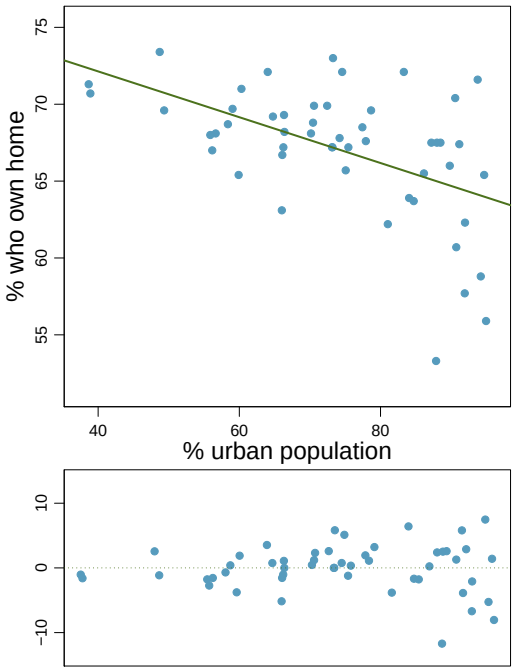
(e) You meet a married man from Britain who is 5'9" (69 inches). What would you predict his wife's height to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict his wife's height? Why or why not?

5.32 Urban homeowners, Part II. Exercise 5.25 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?



5.33 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

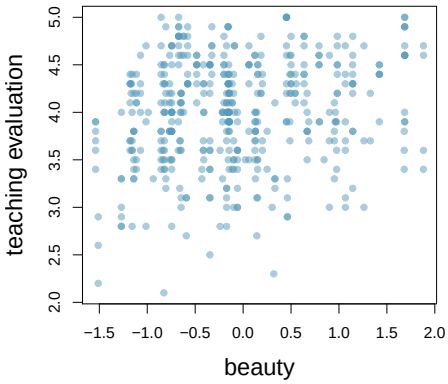
$$\widehat{head\_circumference} = 3.91 + 0.78 \times gestational\_age$$

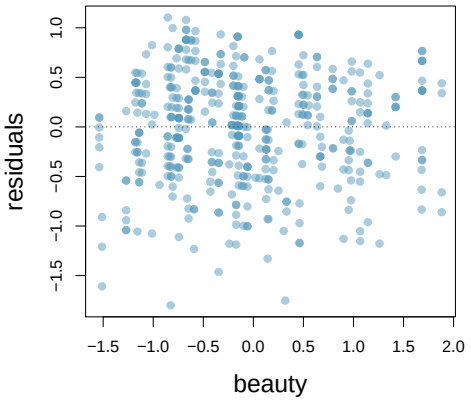
(a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?

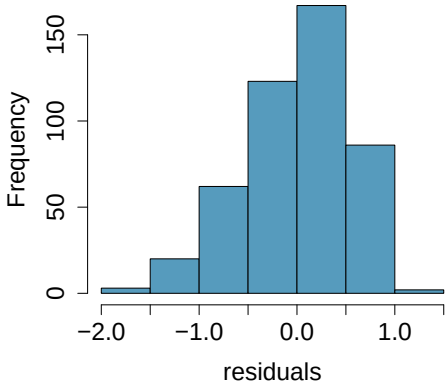
(b) The standard error for the coefficient of gestational age is 0.35, which is associated with $df = 23$. Does the model provide strong evidence that gestational age is significantly associated with head circumference?

5.34 Rate my professor. Some college students critique professors' teaching at RateMyProfessors.com, a web page where students anonymously rate their professors on quality, easiness, and attractiveness. Using the self-selected data from this public forum, researchers examine the relations between quality, easiness, and attractiveness for professors at various universities. In this exercise we will work with a portion of these data that the researchers made publicly available.[23]

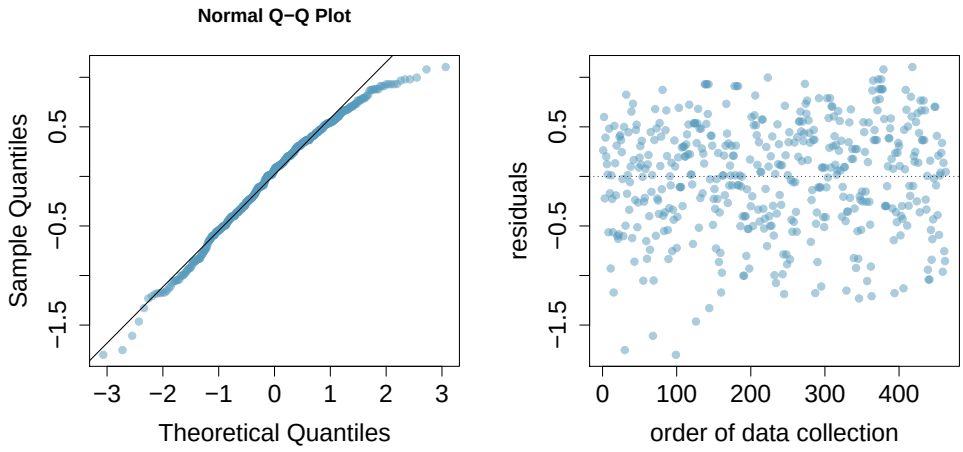
The scatterplot on the right shows the relationship between teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. Given below are associated diagnostic plots. Also given is a regression output for predicting teaching evaluation score from beauty score.







[23] J. Felton et al. "Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness". In: *Assessment & Evaluation in Higher Education* 29.1 (2004), pp. 91–108.



	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.

Chapter 6

Multiple and logistic regression

The principles of simple linear regression lay the foundation for more sophisticated regression methods used in a wide range of challenging settings. In Chapter 6, we explore multiple regression, which introduces the possibility of more than one predictor, and logistic regression, a technique for predicting categorical outcomes with two possible categories.

6.1 Introduction to multiple regression

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider Ebay auctions of a video game called *Mario Kart* for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables. For instance, all other characteristics held constant, are longer auctions associated with higher or lower prices? And, on average, how much more do buyers tend to pay for additional Wii wheels (plastic steering wheels that attach to the Wii controller) in auctions? Multiple regression will help us answer these and other questions.

The data set `mario_kart` includes results from 141 auctions.[1] Four observations from this data set are shown in Table 6.1, and descriptions for each variable are shown in Table 6.2. Notice that the condition and stock photo variables are indicator variables. For instance, the `cond_new` variable takes value 1 if the game up for auction is new and 0 if it is used. Using indicator variables in place of category names allows for these variables to be directly used in regression. See Section 5.2.6 for additional details. Multiple regression also allows for categorical variables with many levels, though we do not have any such variables in this analysis, and we save these details for a second or third course.

[1] Diez DM, Barr CD, and Çetinkaya-Rundel M. 2012. *openintro*: OpenIntro data sets and supplemental functions. cran.r-project.org/web/packages/openintro.

	price	cond_new	stock_photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

Table 6.1: Four observations from the `mario_kart` data set.

variable	description
`price`	final auction price plus shipping costs, in US dollars
`cond_new`	a coded two-level categorical variable, which takes value `1` when the game is new and `0` if the game is used
`stock_photo`	a coded two-level categorical variable, which takes value `1` if the primary photo used in the auction was a stock photo and `0` if the photo was unique to that auction
`duration`	the length of the auction, in days, taking values from 1 to 10
`wheels`	the number of Wii wheels included with the auction (a *Wii wheel* is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart)

Table 6.2: Variables and their descriptions for the `mario_kart` data set.

6.1.1 A single-variable model for the Mario Kart data

Let's fit a linear regression model with the game's condition as a predictor of auction price. The model may be written as

$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

Results of this model are shown in Table 6.3 and a scatterplot for price versus game condition is shown in Figure 6.4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	0.0000
cond_new	10.8996	1.2583	8.66	0.0000
				$df = 139$

Table 6.3: Summary of a linear model for predicting auction price based on game condition.

⊙ **Guided Practice 6.1** Examine Figure 6.4. Does the linear model seem reasonable?[2]

[2]Yes. Constant variability, nearly normal residuals, and linearity all appear reasonable.

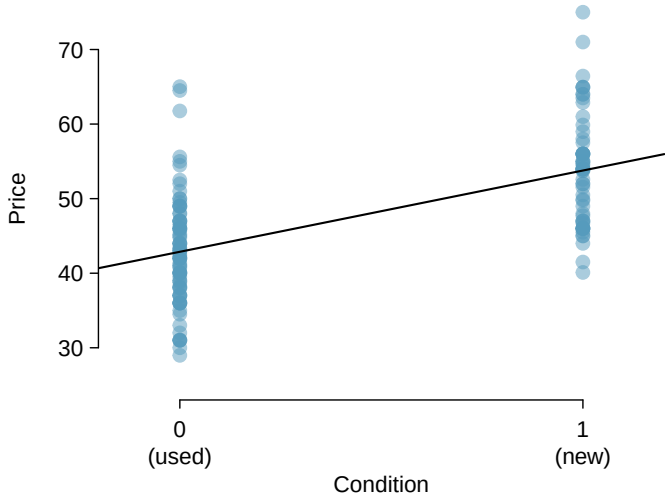


Figure 6.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

Example 6.2 Interpret the coefficient for the game's condition in the model. Is this coefficient significantly different from 0?

Note that `cond_new` is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. So 10.90 means that the model predicts an extra $10.90 for those games that are new versus those that are used. (See Section 5.2.6 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Table 6.3, we can see that the p-value for `cond_new` is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one-variable model.

6.1.2 Including and assessing many variables in a model

Sometimes there are underlying structures or relationships between predictor variables. For instance, new games sold on Ebay tend to come with more Wii wheels, which may have led to higher prices for those auctions. We would like to fit a model that includes all potentially important variables simultaneously. This would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression, such models are a common first step in providing evidence of a causal connection.

We want to construct a model that accounts for not only the game condition, as in Section 6.1.1, but simultaneously accounts for three other variables: `stock_photo`, `duration`, and `wheels`.

$$\begin{aligned}\widehat{\texttt{price}} &= \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} \\ &\quad + \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels} \\ \hat{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \end{aligned} \tag{6.3}$$

In this equation, y represents the total price, x_1 indicates whether the game is new, x_2 indicates whether a stock photo was used, x_3 is the duration of the auction, and x_4 is the number of Wii wheels included with the game. Just as with the single predictor case, a multiple regression model may be missing important components or it might not precisely represent the relationship between the outcome and the available explanatory variables. While no model is perfect, we wish to explore the possibility that this one may fit the data reasonably well.

We estimate the parameters β_0, β_1, ..., β_4 in the same way as we did in the case of a single predictor. We select b_0, b_1, ..., b_4 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{141}^2 = \sum_{i=1}^{141} e_i^2 = \sum_{i=1}^{141} (y_i - \hat{y}_i)^2 \tag{6.4}$$

Here there are 141 residuals, one for each observation. We typically use a computer to minimize the sum in Equation (6.4) and compute point estimates, as shown in the sample output in Table 6.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
				$df = 136$

Table 6.5: Output for the regression model where `price` is the outcome and `cond_new`, `stock_photo`, `duration`, and `wheels` are the predictors.

Multiple regression model

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors. We often estimate the β_i parameters using a computer.

⊙ **Guided Practice 6.5** Write out the model in Equation (6.3) using the point estimates from Table 6.5. How many predictors are there in this model?[3]

⊙ **Guided Practice 6.6** What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?[4]

[3] $\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$, and there are $k = 4$ predictor variables.

[4] It is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$.

⊙ **Guided Practice 6.7** Compute the residual of the first observation in Table 6.1 on page 262 using the equation identified in Guided Practice 6.5. [5]

● **Example 6.8** We estimated a coefficient for `cond_new` in Section 6.1.1 of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `price` and predictor `cond_new` using simple linear regression, we were unable to control for other variables like the number of Wii wheels included in the auction. That model was biased by the confounding variable `wheels`. When we use both variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).

Example 6.8 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** (pronounced as *co-linear*) when they are correlated, and this collinearity complicates model estimation. While it is impossible to prevent collinearity from arising in observational data, experiments are usually designed to prevent predictors from being collinear.

⊙ **Guided Practice 6.9** The estimated value of the intercept is 36.21, and one might be tempted to make some interpretation of this coefficient, such as, it is the model's predicted price when each of the variables take value zero: the game is used, the primary image is not a stock photo, the auction duration is zero days, and there are no wheels included. Is there any value gained by making this interpretation?[6]

6.1.3 Adjusted R^2 as a better estimate of explained variance

We first used R^2 in Section 5.2 to determine the amount of variability in the response that was explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

where e_i represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can often be even more informative.

⊙ **Guided Practice 6.10** The variance of the residuals for the model given in Guided Practice 6.7 is 23.34, and the variance of the total price in all the auctions is 83.06. Calculate R^2 for this model.[7]

---

[5] $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the variables values from the observation and the equation identified in Guided Practice 6.5.

[6] Three of the variables (`cond_new`, `stock_photo`, and `wheels`) do take value 0, but the auction duration is always one or more days. If the auction is not up for any days, then no one can bid on it! That means the total auction price would always be zero for such an auction; the interpretation of the intercept in this setting is not insightful.

[7] $R^2 = 1 - \frac{23.34}{83.06} = 0.719$.

This strategy for estimating R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is actually a biased estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R^2.

Adjusted R^2 as a tool for model assessment

The **adjusted R^2** is computed as

$$R^2_{adj} = 1 - \frac{Var(e_i)/(n-k-1)}{Var(y_i)/(n-1)} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model.

Because k is never negative, the adjusted R^2 will be smaller – often times just a little smaller – than the unadjusted R^2. The reasoning behind the adjusted R^2 lies in the **degrees of freedom** associated with each variance.[8]

⊙ **Guided Practice 6.11** There were $n = 141$ auctions in the `mario_kart` data set and $k = 4$ predictor variables in the model. Use n, k, and the variances from Guided Practice 6.10 to calculate R^2_{adj} for the Mario Kart model.[9]

⊙ **Guided Practice 6.12** Suppose you added another predictor to the model, but the variance of the errors $Var(e_i)$ didn't go down. What would happen to the R^2? What would happen to the adjusted R^2? [10]

6.2 Model selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate from the model variables that are less important.

In this section, and in practice, the model that includes all available explanatory variables is often referred to as the **full model**. Our goal is to assess whether the full model is the best model. If it isn't, we want to identify a smaller model that is preferable.

[8] In multiple regression, the degrees of freedom associated with the variance of the estimate of the residuals is $n-k-1$, not $n-1$. For instance, if we were to make predictions for new data using our current model, we would find that the unadjusted R^2 is an overly optimistic estimate of the reduction in variance in the response, and using the degrees of freedom in the adjusted R^2 formula helps correct this bias.

[9] $R^2_{adj} = 1 - \frac{23.34}{83.06} \times \frac{141-1}{141-4-1} = 0.711$.

[10] The unadjusted R^2 would stay the same and the adjusted R^2 would go down.

6.2.1 Identifying variables in the model that may not be helpful

Table 6.6 provides a summary of the regression output for the full model for the auction data. The last column of the table lists p-values that can be used to assess hypotheses of the following form:

H_0: $\beta_i = 0$ when the other explanatory variables are included in the model.

H_A: $\beta_i \neq 0$ when the other explanatory variables are included in the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

Table 6.6: The fit for the full regression model, including the adjusted R^2.

Example 6.13 The coefficient of `cond_new` has a t test statistic of $T = 4.88$ and a p-value for its corresponding hypotheses ($H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$) of about zero. How can this be interpreted?

If we keep all the other variables in the model and add no others, then there is strong evidence that a game's condition (new or used) has a real relationship with the total auction price.

Example 6.14 Is there strong evidence that using a stock photo is related to the total auction price?

The t test statistic for `stock_photo` is $T = 1.02$ and the p-value is about 0.31. After accounting for the other predictors, there is not strong evidence that using a stock photo in an auction is related to the total price of the auction. We might consider removing the `stock_photo` variable from the model.

Guided Practice 6.15 Identify the p-values for both the `duration` and `wheels` variables in the model. Is there strong evidence supporting the connection of these variables with the total price in the model?[11]

There is not statistically significant evidence that either the stock photo or duration variables contribute meaningfully to the model. Next we consider common strategies for pruning such variables from a model.

[11] The p-value for the auction duration is 0.8882, which indicates that there is not statistically significant evidence that the duration is related to the total auction price when accounting for the other variables. The p-value for the Wii wheels variable is about zero, indicating that this variable is associated with the total auction price.

TIP: Using adjusted R^2 instead of p-values for model selection
The adjusted R^2 may be used as an alternative to p-values for model selection, where a higher adjusted R^2 represents a better model fit. For instance, we could compare two models using their adjusted R^2, and the model with the higher adjusted R^2 would be preferred. This approach tends to include more variables in the final model when compared to the p-value approach.

6.2.2 Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward-selection* and *forward-selection*. These techniques are often referred to as **stepwise** model selection strategies, because they add or delete one variable at a time as they "step" through the candidate predictors. We will discuss these strategies in the context of the p-value approach. Alternatively, we could have employed an R^2_{adj} approach.

The **backward-elimination** strategy starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until only variables with statistically significant p-values remain. The strategy within each elimination step is to drop the variable with the largest p-value, refit the model, and reassess the inclusion of all variables.

Example 6.16 Results corresponding to the *full model* for the `mario_kart` data are shown in Table 6.6. How should we proceed under the backward-elimination strategy?

There are two variables with coefficients that are not statistically different from zero: `stock_photo` and `duration`. We first drop the `duration` variable since it has a larger corresponding p-value, *then we refit the model*. A regression summary for the new model is shown in Table 6.7.

In the new model, there is not strong evidence that the coefficient for `stock_photo` is different from zero, even though the p-value decreased slightly, and the other p-values remain very small. Next, we again eliminate the variable with the largest non-significant p-value, `stock_photo`, and refit the model. The updated regression summary is shown in Table 6.8.

In the latest model, we see that the two remaining predictors have statistically significant coefficients with p-values of about zero. Since there are no variables remaining that could be eliminated from the model, we stop. The final model includes only the `cond_new` and `wheels` variables in predicting the total auction price:

$$\begin{aligned} \hat{y} &= b_0 + b_1 x_1 + b_4 x_4 \\ &= 36.78 + 5.58 x_1 + 7.23 x_4 \end{aligned}$$

where x_1 represents `cond_new` and x_4 represents `wheels`.

An alternative to using p-values in model selection is to use the adjusted R^2. At each elimination step, we refit the model without each of the variables up for potential elimination. For example, in the first step, we would fit four models, where each would be missing a different predictor. If one of these smaller models has a higher adjusted R^2 than our current model, we pick the smaller model with the largest adjusted R^2. We continue in this way until removing variables does not increase R^2_{adj}. Had we

used the adjusted R^2 criteria, we would have kept the `stock_photo` variable along with the `cond_new` and `wheels` variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.0483	0.9745	36.99	0.0000
cond_new	5.1763	0.9961	5.20	0.0000
stock_photo	1.1177	1.0192	1.10	0.2747
wheels	7.2984	0.5448	13.40	0.0000
$R^2_{adj} = 0.7128$				$df = 137$

Table 6.7: The output for the regression model where `price` is the outcome and the duration variable has been eliminated from the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.7849	0.7066	52.06	0.0000
cond_new	5.5848	0.9245	6.04	0.0000
wheels	7.2328	0.5419	13.35	0.0000
$R^2_{adj} = 0.7124$				$df = 138$

Table 6.8: The output for the regression model where `price` is the outcome and the duration and stock photo variables have been eliminated from the model.

Notice that the p-value for `stock_photo` changed a little from the full model (0.309) to the model that did not include the `duration` variable (0.275). It is common for p-values of one variable to change, due to collinearity, after eliminating a different variable. This fluctuation emphasizes the importance of refitting a model after each variable elimination step. The p-values tend to change dramatically when the eliminated variable is highly correlated with another variable in the model.

The **forward-selection** strategy is the reverse of the backward-elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that present strong evidence of their importance in the model.

Example 6.17 Construct a model for the `mario_kart` data set using the forward-selection strategy.

We start with the model that includes no variables. Then we fit each of the possible models with just one variable. That is, we fit the model including just the `cond_new` predictor, then the model including just the `stock_photo` variable, then a model with just `duration`, and a model with just `wheels`. Each of the four models (yes, we fit four models!) provides a p-value for the coefficient of the predictor variable. Out of these four variables, the `wheels` variable had the smallest p-value. Since its p-value is less than 0.05 (the p-value was smaller than 2e-16), we add the Wii wheels variable to the model. Once a variable is added in forward-selection, it will be included in all models considered as well as the final model.

Since we successfully found a first variable to add, we consider adding another. We fit three new models: (1) the model including just the `cond_new` and `wheels` variables (output in Table 6.8), (2) the model including just the `stock_photo` and `wheels`

variables, and (3) the model including only the `duration` and `wheels` variables. Of these models, the first had the lowest p-value for its new variable (the p-value corresponding to `cond_new` was 1.4e-08). Because this p-value is below 0.05, we add the `cond_new` variable to the model. Now the final model is guaranteed to include both the condition and wheels variables.

We must then repeat the process a third time, fitting two new models: (1) the model including the `stock_photo`, `cond_new`, and `wheels` variables (output in Table 6.7) and (2) the model including the `duration`, `cond_new`, and `wheels` variables. The p-value corresponding to `stock_photo` in the first model (0.275) was smaller than the p-value corresponding to `duration` in the second model (0.682). However, since this smaller p-value was not below 0.05, there was not strong evidence that it should be included in the model. Therefore, neither variable is added and we are finished.

The final model is the same as that arrived at using the backward-selection strategy.

Example 6.18 As before, we could have used the R^2_{adj} criteria instead of examining p-values in selecting variables for the model. Rather than look for variables with the smallest p-value, we look for the model with the largest R^2_{adj}. What would the result of forward-selection be using the adjusted R^2 approach?

Using the forward-selection strategy, we start with the model with no predictors. Next we look at each model with a single predictor. If one of these models has a larger R^2_{adj} than the model with no variables, we use this new model. We repeat this procedure, adding one variable at a time, until we cannot find a model with a larger R^2_{adj}. If we had done the forward-selection strategy using R^2_{adj}, we would have arrived at the model including `cond_new`, `stock_photo`, and `wheels`, which is a slightly larger model than we arrived at using the p-value approach and the same model we arrived at using the adjusted R^2 and backwards-elimination.

Model selection strategies

The backward-elimination strategy begins with the largest model and eliminates variables one-by-one until we are satisfied that all remaining variables are important to the model. The forward-selection strategy starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

There is no guarantee that the backward-elimination and forward-selection strategies will arrive at the same final model using the p-value or adjusted R^2 methods. If the backwards-elimination and forward-selection strategies are both tried and they arrive at different models, choose the model with the larger R^2_{adj} as a tie-breaker; other tie-break options exist but are beyond the scope of this book.

It is generally acceptable to use just one strategy, usually backward-elimination with either the p-value or adjusted R^2 criteria. However, before reporting the model results, we must verify the model conditions are reasonable.

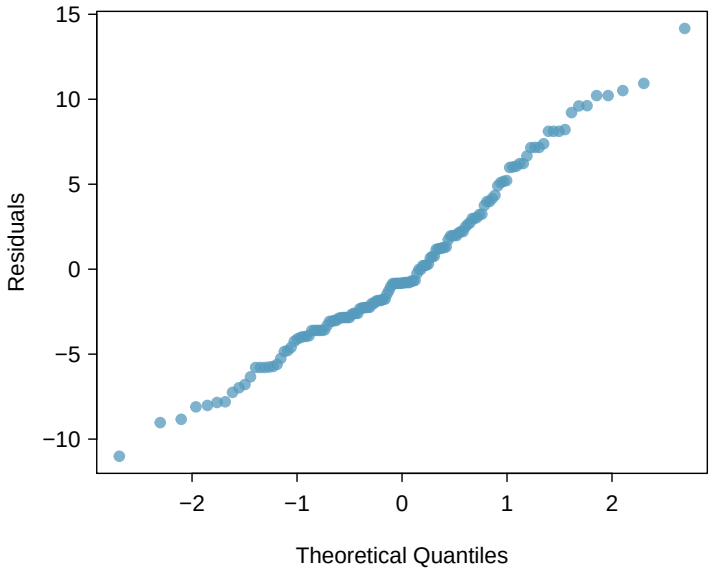


Figure 6.9: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

6.3 Checking model assumptions using graphs

Multiple regression methods using the model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

generally depend on the following four assumptions:

1. the residuals of the model are nearly normal,
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

Simple and effective plots can be used to check each of these assumptions. We will consider the model for the auction data that uses the game condition and number of wheels as predictors. The plotting methods presented here may also be used to check the conditions for the models introduced in Chapter 5.

Normal probability plot. A normal probability plot of the residuals is shown in Figure 6.9. While the plot exhibits some minor irregularities, there are no outliers that might be cause for concern. In a normal probability plot for residuals, we tend to be most worried about residuals that appear to be outliers, since these indicate long tails in the distribution of residuals.

Absolute values of residuals against fitted values. A plot of the absolute value of the residuals against their corresponding fitted values ($\hat{y}_i$) is shown in Figure 6.10. This plot is helpful to check the condition that the variance of the residuals is approximately constant. We don't see any obvious deviations from constant variance in this example.

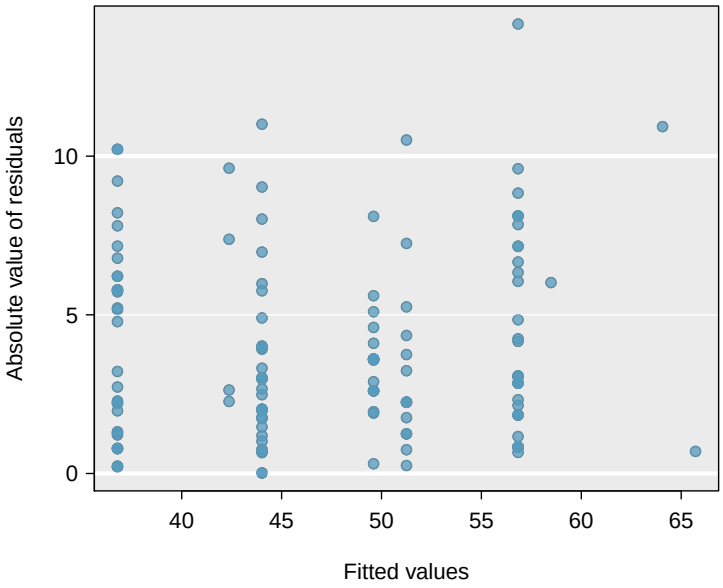


Figure 6.10: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

Residuals in order of their data collection. A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 6.11. Such a plot is helpful in identifying any connection between cases that are close to one another, e.g. we could look for declining prices over time or if there was a time of the day when auctions tended to fetch a higher price. Here we see no structure that indicates a problem.[12]

Residuals against each predictor variable. We consider a plot of the residuals against the `cond_new` variable and the residuals against the `wheels` variable. These plots are shown in Figure 6.12. For the two-level condition variable, we are guaranteed not to see any remaining trend, and instead we are checking that the variability doesn't fluctuate across groups. In this example, when we consider the residuals against the `wheels` variable, we see some possible structure. There appears to be curvature in the residuals, indicating the relationship is probably not linear.

It is necessary to summarize diagnostics for any model fit. If the diagnostics support the model assumptions, this would improve credibility in the findings. If the diagnostic assessment shows remaining underlying structure in the residuals, we should try to adjust the model to account for that structure. If we are unable to do so, we may still report the model but must also note its shortcomings. In the case of the auction data, we report that there may be a nonlinear relationship between the total price and the number of wheels included for an auction. This information would be important to buyers and sellers; omitting this information could be a setback to the very people who the model might assist.

[12] An especially rigorous check would use **time series** methods. For instance, we could check whether consecutive residuals are correlated. Doing so with these residuals yields no statistically significant correlations.

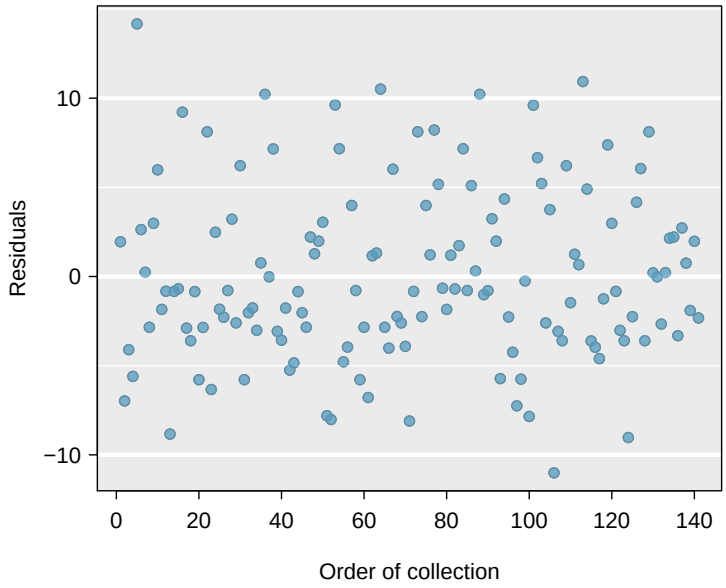


Figure 6.11: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

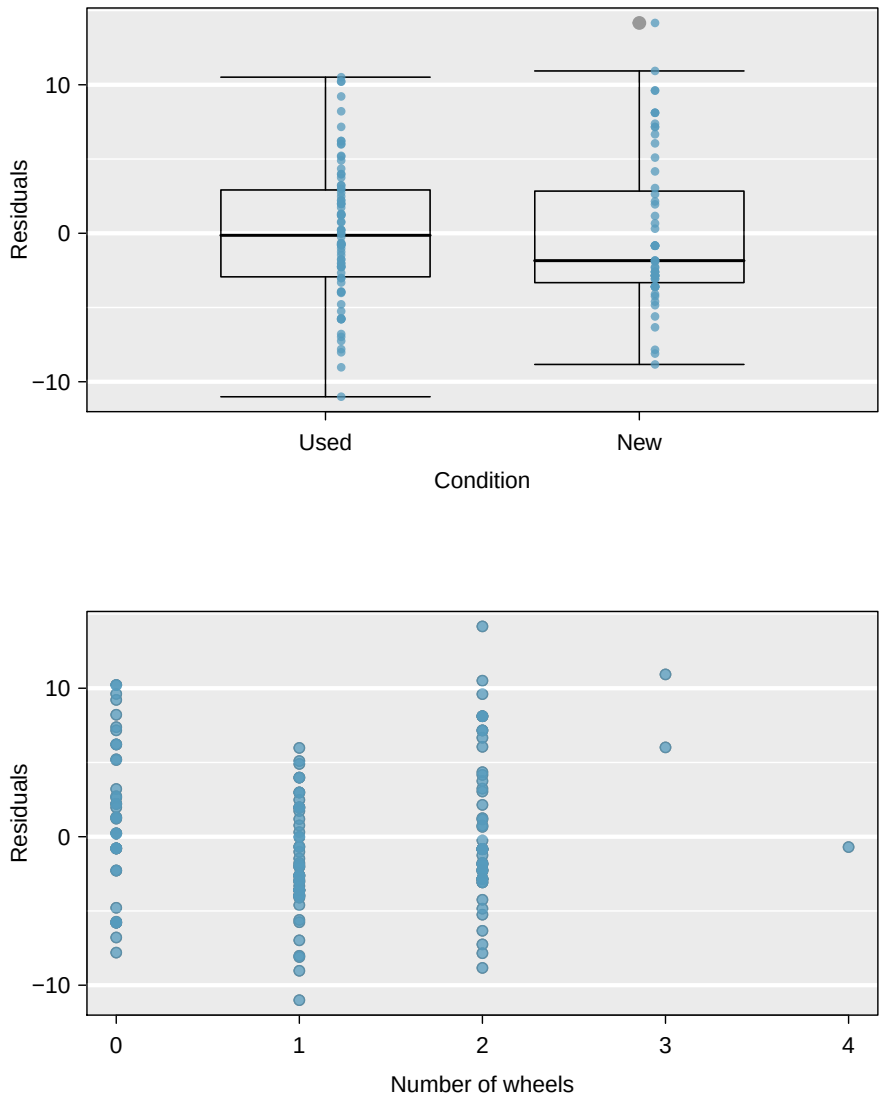


Figure 6.12: In the two-level variable for the game's condition, we check for differences in distribution shape or variability. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable.

"All models are wrong, but some are useful" -George E.P. Box
The truth is that no model is perfect. However, even imperfect models can be useful. Reporting a flawed model can be reasonable so long as we are clear and report the model's shortcomings.

Caution: Don't report results when assumptions are grossly violated
While there is a little leeway in model assumptions, don't go too far. If model assumptions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help.

TIP: Confidence intervals in multiple regression
Confidence intervals for coefficients in multiple regression can be computed using the same formula as in the single predictor model:

$$b_i \pm t^\star_{df} SE_{b_i}$$

where $t^\star_{df}$ is the appropriate t value corresponding to the confidence level and model degrees of freedom, $df = n - k - 1$.

6.4 Logistic regression

In this section we introduce **logistic regression** as a tool for building models when there is a categorical response variable with two levels. Logistic regression is a type of **generalized linear model** (GLM) for response variables where regular multiple regression does not work very well. In particular, the response variable in these settings often takes a form where residuals look completely different from the normal distribution.

GLMs can be thought of as a two-stage modeling approach. We first model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression.

In Section 6.4 we will revisit the `email` data set from Chapter 1. These emails were collected from a single email account, and we will work on developing a basic spam filter using these data. The response variable, `spam`, has been encoded to take value 0 when a message is not spam and 1 when it is spam. Our task will be to build an appropriate model that classifies messages as spam or not spam using email characteristics coded as predictor variables. While this model will not be the same as those used in large-scale spam filters, it shares many of the same features.

variable	description
spam	Specifies whether the message was spam.
to_multiple	An indicator variable for if more than one person was listed in the *To* field of the email.
cc	An indicator for if someone was CCed on the email.
attach	An indicator for if there was an attachment, such as a document or image.
dollar	An indicator for if the word "dollar" or dollar symbol ($) appeared in the email.
winner	An indicator for if the word "winner" appeared in the email message.
inherit	An indicator for if the word "inherit" (or a variation, like "inheritance") appeared in the email.
password	An indicator for if the word "password" was present in the email.
format	Indicates if the email contained special formatting, such as bolding, tables, or links
re_subj	Indicates whether "Re:" was included at the start of the email subject.
exclaim_subj	Indicates whether any exclamation point was included in the email subject.

Table 6.13: Descriptions for 11 variables in the email data set. Notice that all of the variables are indicator variables, which take the value 1 if the specified characteristic is present and 0 otherwise.

6.4.1 Email data

The email data set was first presented in Chapter 1 with a relatively small number of variables. In fact, there are many more variables available that might be useful for classifying spam. Descriptions of these variables are presented in Table 6.13. The spam variable will be the outcome, and the other 10 variables will be the model predictors. While we have limited the predictors used in this section to be categorical variables (where many are represented as indicator variables), numerical predictors may also be used in logistic regression. See the footnote for an additional discussion on this topic.[13]

6.4.2 Modeling the probability of an event

TIP: Notation for a logistic regression model

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the email application, Y_i will be used to represent whether email i is spam ($Y_i = 1$) or not ($Y_i = 0$).

The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation i, $x_{2,i}$ is the value of variable 2 for observation i, and so on.

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 (in our application, this represents a spam message) with probability p_i and the value 0 with probability $1 - p_i$. It is the probability p_i that we model in relation to the predictor variables.

[13] Recall from Chapter 5 that if outliers are present in predictor variables, the corresponding observations may be especially influential on the resulting model. This is the motivation for omitting the numerical variables, such as the number of characters and line breaks in emails, that we saw in Chapter 1. These variables exhibited extreme skew. We could resolve this issue by transforming these variables (e.g. using a log-transformation), but we will omit this further investigation for brevity.

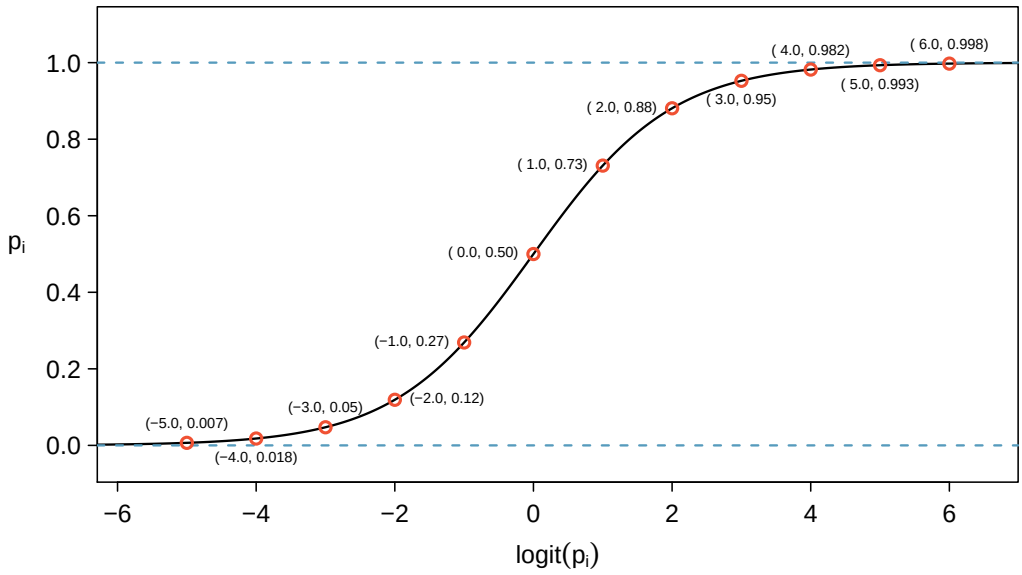


Figure 6.14: Values of p_i against values of $logit(p_i)$.

The logistic regression model relates the probability an email is spam (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots \beta_k x_{k,i} \tag{6.19}$$

We want to choose a transformation in Equation (6.19) that makes practical and mathematical sense. For example, we want a transformation that makes the range of possibilities on the left hand side of Equation (6.19) equal to the range of possibilities for the right hand side; if there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range. A common transformation for p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e \left(\frac{p_i}{1 - p_i} \right)$$

The logit transformation is shown in Figure 6.14. Below, we rewrite Equation (6.19) using the logit transformation of p_i:

$$\log_e \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

In our spam example, there are 10 predictor variables, so $k = 10$. This model isn't very intuitive, but it still has some resemblance to multiple regression, and we can fit this model using software. In fact, once we look at results from software, it will start to feel like we're back in multiple regression, even if the interpretation of the coefficients is more complex.

Example 6.20 Here we create a spam filter with a single predictor: `to_multiple`. This variable indicates whether more than one email address was listed in the *To* field of the email. The following logistic regression model was fit using statistical software:

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.81 \times \texttt{to\_multiple}$$

If an email is randomly selected and it has just one address in the *To* field, what is the probability it is spam? What if more than one address is listed in the *To* field?

If there is only one email in the *To* field, then `to_multiple` takes value 0 and the right side of the model equation equals -2.12. Solving for p_i: $\frac{e^{-2.12}}{1+e^{-2.12}} = 0.11$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we will do the same for this probability: $\hat{p}_i = 0.11$.

If there is more than one address listed in the *To* field, then the right side of the model equation is $-2.12 - 1.81 \times 1 = -3.93$, which corresponds to a probability $\hat{p}_i = 0.02$.

Notice that we could examine -2.12 and -3.93 in Figure 6.14 to estimate the probability before formally calculating the value.

To convert from values on the regression-scale (e.g. -2.12 and -3.93 in Example 6.20), use the following formula, which is the result of solving for p_i in the regression model:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 \ + \ e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute the point estimates for the parameters (the β_i) so that we may make use of this formula. In Example 6.20, the probabilities were calculated as

$$\frac{e^{-2.12}}{1 \ + \ e^{-2.12}} = 0.11 \qquad\qquad \frac{e^{-2.12-1.81}}{1 \ + \ e^{-2.12-1.81}} = 0.02$$

While the information about whether the email is addressed to multiple people is a helpful start in classifying email as spam or not, the probabilities of 11% and 2% are not dramatically different, and neither provides very strong evidence about which particular email messages are spam. To get more precise estimates, we'll need to include many more variables in the model.

We used statistical software to fit the logistic regression model with all ten predictors described in Table 6.13. Like multiple regression, the result may be presented in a summary table, which is shown in Table 6.15. The structure of this table is almost identical to that of multiple regression; the only notable difference is that the p-values are calculated using the normal distribution rather than the t distribution.

Just like multiple regression, we could trim some variables from the model using the p-value. Using backwards elimination with a p-value cutoff of 0.05 (start with the full model and trim the predictors with p-values greater than 0.05), we ultimately eliminate the `exclaim_subj`, `dollar`, `inherit`, and `cc` predictors. The remainder of this section will rely on this smaller model, which is summarized in Table 6.16.

Guided Practice 6.21 Examine the summary of the reduced model in Table 6.16, and in particular, examine the `to_multiple` row. Is the point estimate the same as we found before, -1.81, or is it different? Explain why this might be.[14]

[14] The new estimate is different: -2.87. This new value represents the estimated coefficient when we are also accounting for other variables in the logistic regression model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8362	0.0962	-8.69	0.0000
to_multiple	-2.8836	0.3121	-9.24	0.0000
winner	1.7038	0.3254	5.24	0.0000
format	-1.5902	0.1239	-12.84	0.0000
re_subj	-2.9082	0.3708	-7.84	0.0000
exclaim_subj	0.1355	0.2268	0.60	0.5503
cc	-0.4863	0.3054	-1.59	0.1113
attach	0.9790	0.2170	4.51	0.0000
dollar	-0.0582	0.1589	-0.37	0.7144
inherit	0.2093	0.3197	0.65	0.5127
password	-1.4929	0.5295	-2.82	0.0048

Table 6.15: Summary table for the full logistic regression model for the spam filter example.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8595	0.0910	-9.44	0.0000
to_multiple	-2.8372	0.3092	-9.18	0.0000
winner	1.7370	0.3218	5.40	0.0000
format	-1.5569	0.1207	-12.90	0.0000
re_subj	-3.0482	0.3630	-8.40	0.0000
attach	0.8643	0.2042	4.23	0.0000
password	-1.4871	0.5290	-2.81	0.0049

Table 6.16: Summary table for the logistic regression model for the spam filter, where variable selection has been performed.

Point estimates will generally change a little – and sometimes a lot – depending on which other variables are included in the model. This is usually due to colinearity in the predictor variables. We previously saw this in the Ebay auction example when we compared the coefficient of `cond_new` in a single-variable model and the corresponding coefficient in the multiple regression model that used three additional variables (see Sections 6.1.1 and 6.1.2).

Example 6.22 Spam filters are built to be automated, meaning a piece of software is written to collect information about emails as they arrive, and this information is put in the form of variables. These variables are then put into an algorithm that uses a statistical model, like the one we've fit, to classify the email. Suppose we write software for a spam filter using the reduced model shown in Table 6.16. If an incoming email has the word "winner" in it, will this raise or lower the model's calculated probability that the incoming email is spam?

The estimated coefficient of `winner` is positive (1.7370). A positive coefficient estimate in logistic regression, just like in multiple regression, corresponds to a positive association between the predictor and response variables when accounting for the other variables in the model. Since the response variable takes value 1 if an email is spam and 0 otherwise, the positive coefficient indicates that the presence of "winner" in an email raises the model probability that the message is spam.

Example 6.23 Suppose the same email from Example 6.22 was in HTML format, meaning the `format` variable took value 1. Does this characteristic increase or decrease the probability that the email is spam according to the model?

Since HTML corresponds to a value of 1 in the `format` variable and the coefficient of this variable is negative (-1.5569), this would lower the probability estimate returned from the model.

6.4.3 Practical decisions in the email application

Examples 6.22 and 6.23 highlight a key feature of logistic and multiple regression. In the spam filter example, some email characteristics will push an email's classification in the direction of spam while other characteristics will push it in the opposite direction.

If we were to implement a spam filter using the model we have fit, then each future email we analyze would fall into one of three categories based on the email's characteristics:

1. The email characteristics generally indicate the email is not spam, and so the resulting probability that the email is spam is quite low, say, under 0.05.

2. The characteristics generally indicate the email is spam, and so the resulting probability that the email is spam is quite large, say, over 0.95.

3. The characteristics roughly balance each other out in terms of evidence for and against the message being classified as spam. Its probability falls in the remaining range, meaning the email cannot be adequately classified as spam or not spam.

If we were managing an email service, we would have to think about what should be done in each of these three instances. In an email application, there are usually just two possibilities: filter the email out from the regular inbox and put it in a "spambox", or let the email go to the regular inbox.

Guided Practice 6.24 The first and second scenarios are intuitive. If the evidence strongly suggests a message is not spam, send it to the inbox. If the evidence strongly suggests the message is spam, send it to the spambox. How should we handle emails in the third category?[15]

Guided Practice 6.25 Suppose we apply the logistic model we have built as a spam filter and that 100 messages are placed in the spambox over 3 months. If we used the guidelines above for putting messages into the spambox, about how many legitimate (non-spam) messages would you expect to find among the 100 messages?[16]

Almost any classifier will have some error. In the spam filter guidelines above, we have decided that it is okay to allow up to 5% of the messages in the spambox to be real messages. If we wanted to make it a little harder to classify messages as spam, we could use a cutoff of 0.99. This would have two effects. Because it raises the standard for what can be classified as spam, it reduces the number of good emails that are classified as spam.

[15] In this particular application, we should err on the side of sending more mail to the inbox rather than mistakenly putting good messages in the spambox. So, in summary: emails in the first and last categories go to the regular inbox, and those in the second scenario go to the spambox.

[16] First, note that we proposed a cutoff for the predicted probability of 0.95 for spam. In a worst case scenario, all the messages in the spambox had the minimum probability equal to about 0.95. Thus, we should expect to find about 5 or fewer legitimate messages among the 100 messages placed in the spambox.

However, it will also fail to correctly classify an increased fraction of spam messages. No matter the complexity and the confidence we might have in our model, these practical considerations are absolutely crucial to making a helpful spam filter. Without them, we could actually do more harm than good by using our statistical model.

6.4.4 Diagnostics for the email classifier

Logistic regression conditions

There are two key conditions for fitting a logistic regression model:

1. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.
2. Each outcome Y_i is independent of the other outcomes.

The first condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 3,921 emails in our data set! Let's first visualize these data by plotting the true classification of the emails against the model's fitted probabilities, as shown in Figure 6.17. The vast majority of emails (spam or not) still have fitted probabilities below 0.5.

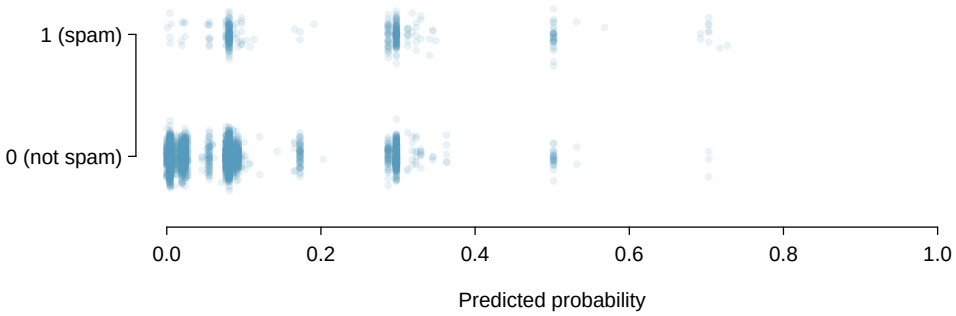


Figure 6.17: The predicted probability that each of the 3,912 emails is spam is classified by their grouping, spam or not. Noise (small, random vertical shifts) have been added to each point so that points with nearly identical values aren't plotted exactly on top of one another. This makes it possible to see more observations.

This may at first seem very discouraging: we have fit a logistic model to create a spam filter, but no emails have a fitted probability of being spam above 0.75. Don't despair; we will discuss ways to improve the model through the use of better variables in Section 6.4.5.

We'd like to assess the quality of our model. For example, we might ask: if we look at emails that we modeled as having a 10% chance of being spam, do we find about 10% of them actually are spam? To help us out, we'll borrow an advanced statistical method called **natural splines** that estimates the local probability over the region 0.00 to 0.75 (the largest predicted probability was 0.73, so we avoid extrapolating). All you need to know about natural splines to understand what we are doing is that they are used to fit flexible lines rather than straight lines.

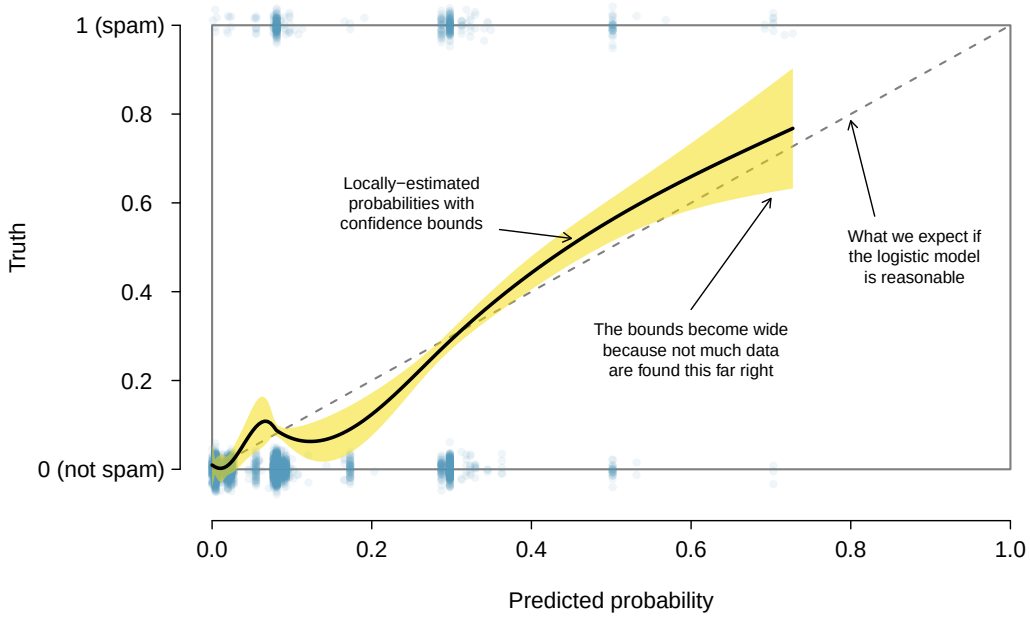


Figure 6.18: The solid black line provides the empirical estimate of the probability for observations based on their predicted probabilities (confidence bounds are also shown for this line), which is fit using natural splines. A small amount of noise was added to the observations in the plot to allow more observations to be seen.

The curve fit using natural splines is shown in Figure 6.18 as a solid black line. If the logistic model fits well, the curve should closely follow the dashed $y = x$ line. We have added shading to represent the confidence bound for the curved line to clarify what fluctuations might plausibly be due to chance. Even with this confidence bound, there are weaknesses in the first model assumption. The solid curve and its confidence bound dips below the dashed line from about 0.1 to 0.3, and then it drifts above the dashed line from about 0.35 to 0.55. These deviations indicate the model relating the parameter to the predictors does not closely resemble the true relationship.

We could evaluate the second logistic regression model assumption – independence of the outcomes – using the model residuals. The residuals for a logistic regression model are calculated the same way as with multiple regression: the observed outcome minus the expected outcome. For logistic regression, the expected value of the outcome is the fitted probability for the observation, and the residual may be written as

$$e_i = Y_i - \hat{p}_i$$

We could plot these residuals against a variety of variables or in their order of collection, as we did with the residuals in multiple regression. However, since the model will need to be revised to effectively classify spam and you have already seen similar residual plots in Section 6.3, we won't investigate the residuals here.

6.4.5 Improving the set of variables for a spam filter

If we were building a spam filter for an email service that managed many accounts (e.g. Gmail or Hotmail), we would spend much more time thinking about additional variables that could be useful in classifying emails as spam or not. We also would use transformations or other techniques that would help us include strongly skewed numerical variables as predictors.

Take a few minutes to think about additional variables that might be useful in identifying spam. Below is a list of variables we think might be useful:

(1) An indicator variable could be used to represent whether there was prior two-way correspondence with a message's sender. For instance, if you sent a message to john@example.com and then John sent you an email, this variable would take value 1 for the email that John sent. If you had never sent John an email, then the variable would be set to 0.

(2) A second indicator variable could utilize an account's past spam flagging information. The variable could take value 1 if the sender of the message has previously sent messages flagged as spam.

(3) A third indicator variable could flag emails that contain links included in previous spam messages. If such a link is found, then set the variable to 1 for the email. Otherwise, set it to 0.

The variables described above take one of two approaches. Variable (1) is specially designed to capitalize on the fact that spam is rarely sent between individuals that have two-way communication. Variables (2) and (3) are specially designed to flag common spammers or spam messages. While we would have to verify using the data that each of the variables is effective, these seem like promising ideas.

Table 6.19 shows a contingency table for spam and also for the new variable described in (1) above. If we look at the 1,090 emails where there was correspondence with the sender in the preceding 30 days, not one of these message was spam. This suggests variable (1) would be very effective at accurately classifying some messages as not spam. With this single variable, we would be able to send about 28% of messages through to the inbox with confidence that almost none are spam.

	prior correspondence		
	`no`	`yes`	Total
`spam`	367	0	367
`not spam`	2464	1090	3554
Total	2831	1090	3921

Table 6.19: A contingency table for `spam` and a new variable that represents whether there had been correspondence with the sender in the preceding 30 days.

The variables described in (2) and (3) would provide an excellent foundation for distinguishing messages coming from known spammers or messages that take a known form of spam. To utilize these variables, we would need to build databases: one holding email addresses of known spammers, and one holding URLs found in known spam messages. Our access to such information is limited, so we cannot implement these two variables in this

textbook. However, if we were hired by an email service to build a spam filter, these would be important next steps.

In addition to finding more and better predictors, we would need to create a customized logistic regression model for each email account. This may sound like an intimidating task, but its complexity is not as daunting as it may at first seem. We'll save the details for a statistics course where computer programming plays a more central role.

For what is the extremely challenging task of classifying spam messages, we have made a lot of progress. We have seen that simple email variables, such as the format, inclusion of certain words, and other circumstantial characteristics, provide helpful information for spam classification. Many challenges remain, from better understanding logistic regression to carrying out the necessary computer programming, but completing such a task is very nearly within your reach.

6.5 Exercises

6.5.1 Introduction to multiple regression

6.1 Baby weights, Part I. The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.[17]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

(c) Is there a statistically significant relationship between the average birth weight and smoking?

6.2 Baby weights, Part II. Exercise 6.1 introduces a data set on birth weight of babies. Another variable we consider is `parity`, which is 0 if the child is the first born, and 1 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from `parity`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

(c) Is there a statistically significant relationship between the average birth weight and parity?

[17] Child Health and Development Studies, Baby weights data set.

6.3 Baby weights, Part III. We considered the variables `smoke` and `parity`, one at a time, in modeling birth weights of babies in Exercises 6.1 and 6.2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

(a) Write the equation of the regression line that includes all of the variables.

(b) Interpret the slopes of `gestation` and `age` in this context.

(c) The coefficient for `parity` is different than in the linear model shown in Exercise 6.2. Why might there be a difference?

(d) Calculate the residual for the first observation in the data set.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

6.4 Absenteeism. Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

	eth	sex	lrn	days
1	0	1	1	2
2	0	1	1	11
⋮	⋮	⋮	⋮	⋮
146	1	0	0	37

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on ethnic background (`eth`: 0 - aboriginal, 1 - not aboriginal), sex (`sex`: 0 - female, 1 - male), and learner status (`lrn`: 0 - average learner, 1 - slow learner).[18]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.93	2.57	7.37	0.0000
eth	-9.11	2.60	-3.51	0.0000
sex	3.10	2.64	1.18	0.2411
lrn	2.15	2.65	0.81	0.4177

(a) Write the equation of the regression line.

(b) Interpret each one of the slopes in this context.

(c) Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.

(d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R^2 and the adjusted R^2. Note that there are 146 observations in the data set.

6.5 GPA. A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown below. Note that male is coded as 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.45	0.35	9.85	0.00
studyweek	0.00	0.00	0.27	0.79
sleepnight	0.01	0.05	0.11	0.91
outnight	0.05	0.05	1.01	0.32
gender	-0.08	0.12	-0.68	0.50

(a) Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.

(b) Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain

[18] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. Data can also be found in the R MASS package. New York: Springer, 2002.

6.6 Cherry trees. Timber yield is approximately equal to the volume of a tree, however, this value is difficult to measure without first cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree's volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height is measured in feet, diameter in inches (at 54 inches above ground), and volume in cubic feet.[19]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.99	8.64	-6.71	0.00
height	0.34	0.13	2.61	0.01
diameter	4.71	0.26	17.82	0.00

(a) Calculate a 95% confidence interval for the coefficient of height, and interpret it in the context of the data.

(b) One tree in this sample is 79 feet tall, has a diameter of 11.3 inches, and is 24.2 cubic feet in volume. Determine if the model overestimates or underestimates the volume of this tree, and by how much.

6.5.2 Model selection

6.7 Baby weights, Part IV. Exercise 6.3 considers a model that predicts a newborn's weight using several predictors. Use the regression table below, which summarizes the model, to answer the following questions. If necessary, refer back to Exercise 6.3 for a reminder about the meaning of each variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

(a) Determine which variables, if any, do not have a significant linear relationship with the outcome and should be candidates for removal from the model. If there is more than one such variable, indicate which one should be removed first.

(b) The summary table below shows the results of the model with the `age` variable removed. Determine if any other variable(s) should be removed from the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.64	14.04	-5.74	0.0000
gestation	0.44	0.03	15.28	0.0000
parity	-3.29	1.06	-3.10	0.0020
height	1.15	0.20	5.64	0.0000
weight	0.05	0.03	2.00	0.0459
smoke	-8.38	0.95	-8.82	0.0000

[19] D.J. Hand. *A handbook of small data sets.* Chapman & Hall/CRC, 1994.

6.8 Absenteeism, Part II. Exercise 6.4 considers a model that predicts the number of days absent using three predictors: ethnic background (`eth`), gender (`sex`), and learner status (`lrn`). Use the regression table below to answer the following questions. If necessary, refer back to Exercise 6.4 for additional details about each variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.93	2.57	7.37	0.0000
eth	-9.11	2.60	-3.51	0.0000
sex	3.10	2.64	1.18	0.2411
lrn	2.15	2.65	0.81	0.4177

(a) Determine which variables, if any, do not have a significant linear relationship with the outcome and should be candidates for removal from the model. If there is more than one such variable, indicate which one should be removed first.

(b) The summary table below shows the results of the regression we refit after removing learner status from the model. Determine if any other variable(s) should be removed from the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.98	2.22	9.01	0.0000
eth	-9.06	2.60	-3.49	0.0006
sex	2.78	2.60	1.07	0.2878

6.9 Baby weights, Part V. Exercise 6.3 provides regression output for the full model (including all explanatory variables available in the data set) for predicting birth weight of babies. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

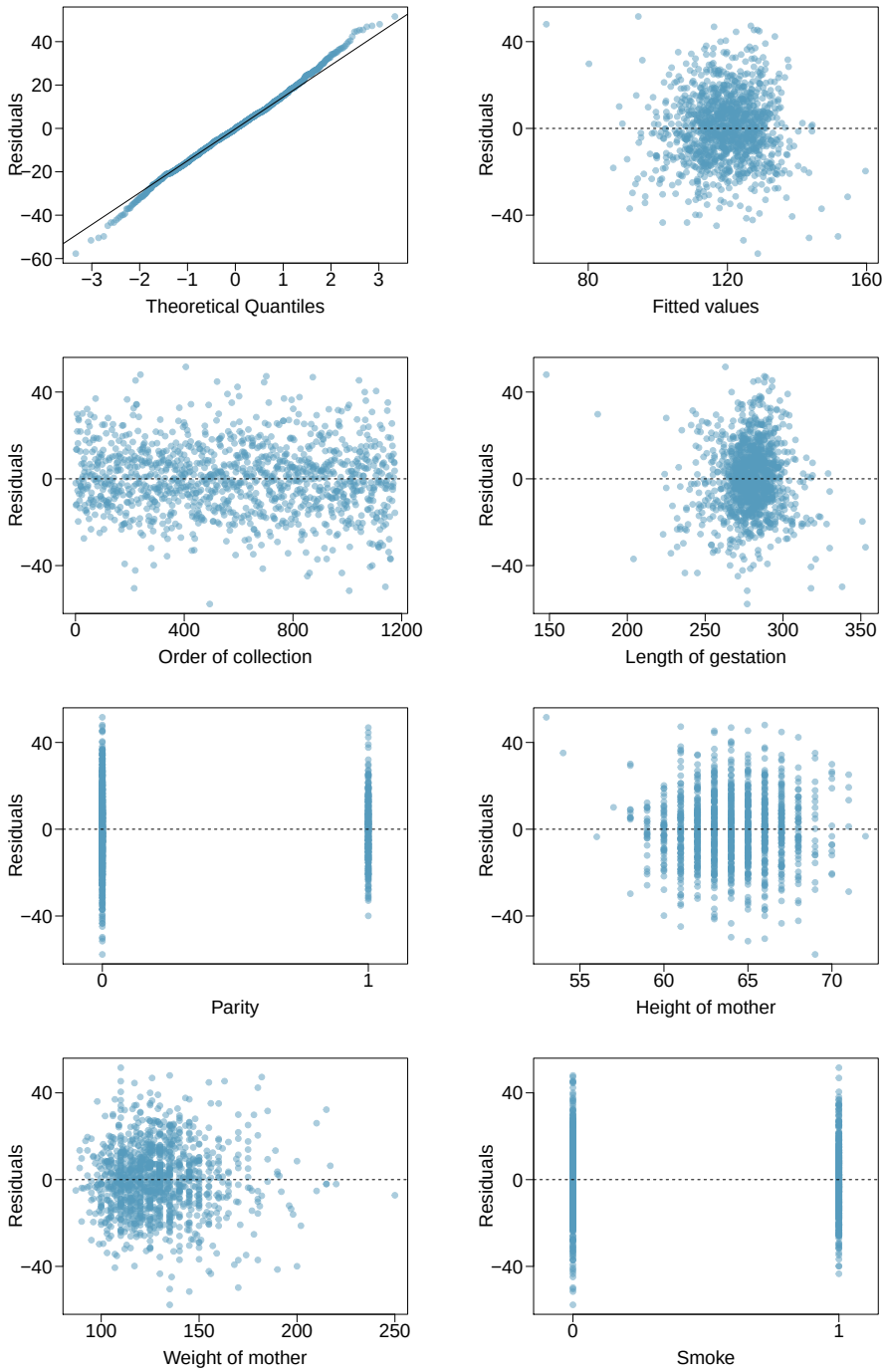
variable	gestation	parity	age	height	weight	smoke
p-value	2.2×10^{-16}	0.1052	0.2375	2.97×10^{-12}	8.2×10^{-8}	2.2×10^{-16}
R^2_{adj}	0.1657	0.0013	0.0003	0.0386	0.0229	0.0569

6.10 Absenteeism, Part III. Exercise 6.4 provides regression output for the full model, including all explanatory variables available in the data set, for predicting the number of days absent from school. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

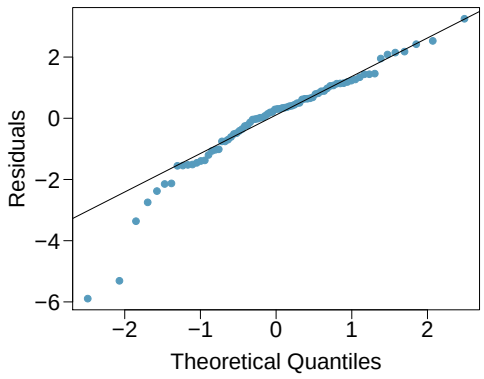
variable	ethnicity	sex	learner status
p-value	0.0007	0.3142	0.5870
R^2_{adj}	0.0714	0.0001	0

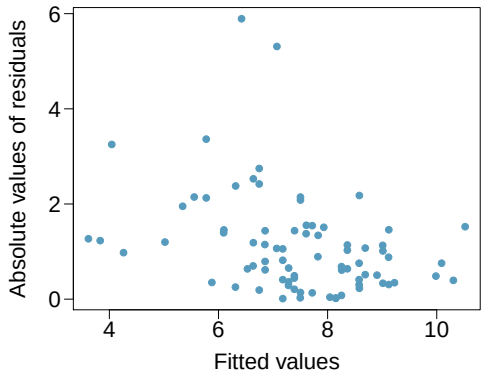
6.5.3 Checking model assumptions using graphs

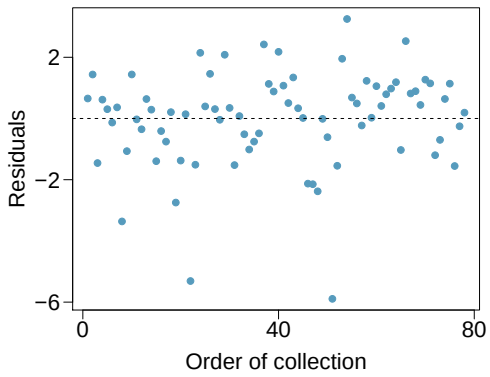
6.11 Baby weights, Part V. Exercise 6.7 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.

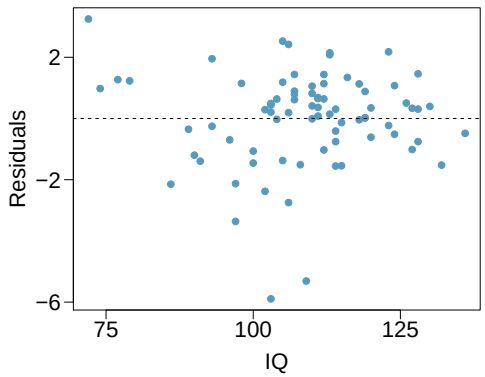


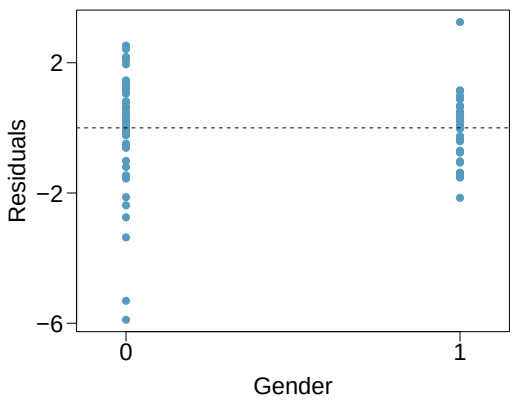
6.12 GPA and IQ. A regression model for predicting GPA from gender and IQ was fit, and both predictors were found to be statistically significant. Using the plots given below, determine if this regression model is appropriate for these data.







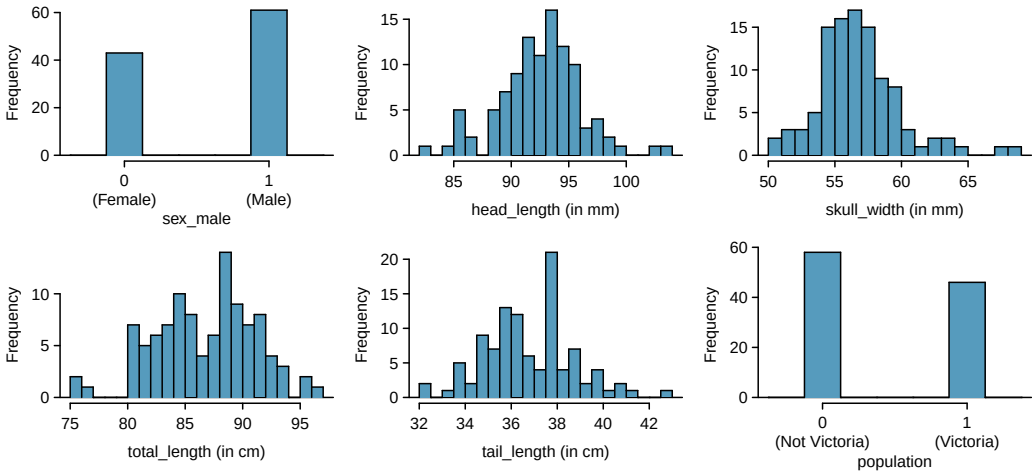




6.5.4 Logistic regression

6.13 Possum classification, Part I. The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum (see Figure 5.5 on page 222). We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia.

We use logistic regression to differentiate between possums in these two regions. The outcome variable, called `population`, takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland. We consider five predictors: `sex_male` (an indicator for a possum being male), `head_length`, `skull_width`, `total_length`, and `tail_length`. Each variable is summarized in a histogram. The full logistic regression model and a reduced model after variable selection are summarized in the table.



	Full Model				*Reduced Model*			
	Estimate	SE	Z	Pr(>|Z|)	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	39.2349	11.5368	3.40	0.0007	33.5095	9.9053	3.38	0.0007
sex_male	-1.2376	0.6662	-1.86	0.0632	-1.4207	0.6457	-2.20	0.0278
head_length	-0.1601	0.1386	-1.16	0.2480				
skull_width	-0.2012	0.1327	-1.52	0.1294	-0.2787	0.1226	-2.27	0.0231
total_length	0.6488	0.1531	4.24	0.0000	0.5687	0.1322	4.30	0.0000
tail_length	-1.8708	0.3741	-5.00	0.0000	-1.8057	0.3599	-5.02	0.0000

(a) Examine each of the predictors. Are there any outliers that are likely to have a very large influence on the logistic regression model?

(b) The summary table for the full model indicates that at least one variable should be eliminated when using the p-value approach for variable selection: `head_length`. The second component of the table summarizes the reduced model following variable selection. Explain why the remaining estimates change between the two models.

6.14 Challenger disaster, Part I. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. *Temp* gives the temperature in Fahrenheit, *Damaged* represents the number of damaged O-rings, and *Undamaged* represents the number of O-rings that were not damaged.

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

(a) Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

(b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below. Describe the key components of this summary table in words.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

(c) Write out the logistic model using the point estimates of the model parameters.

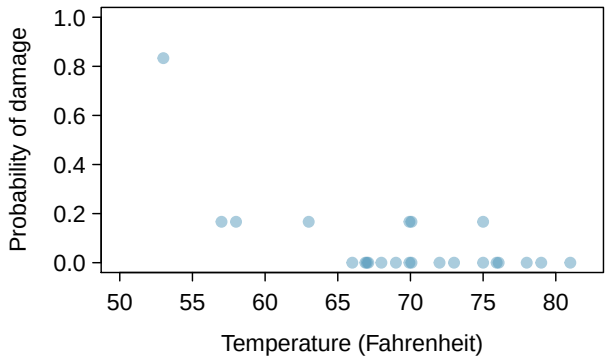
(d) Based on the model, do you think concerns regarding O-rings are justified? Explain.

6.15 Possum classification, Part II. A logistic regression model was proposed for classifying common brushtail possums into their two regions in Exercise 6.13. Use the results of the summary table for the reduced model presented in Exercise 6.13 for the questions below. The outcome variable took value 1 if the possum was from Victoria and 0 otherwise.

(a) Write out the form of the model. Also identify which of the following variables are positively associated (when controlling for other variables) with a possum being from Victoria: `skull_width`, `total_length`, and `tail_length`.

(b) Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the reduced model's computed probability that this possum is from Victoria? How confident are you in the model's accuracy of this probability calculation?

6.16 Challenger disaster, Part II. Exercise 6.14 introduced us to O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoff in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle. See this earlier exercise if you would like to browse the original data.



(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged. Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$$\hat{p}_{57} = 0.341 \qquad \hat{p}_{59} = 0.251 \qquad \hat{p}_{61} = 0.179 \qquad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \qquad \hat{p}_{67} = 0.056 \qquad \hat{p}_{69} = 0.037 \qquad \hat{p}_{71} = 0.024$$

(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

(c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

Appendix A

Probability

Probability forms a foundation for statistics. You might already be familiar with many aspects of probability, however, formalization of the concepts is new for most. This chapter aims to introduce probability on familiar terms using processes most people have seen before.

A.1 Defining probability

Example A.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 A.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 A.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 A.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 A.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.

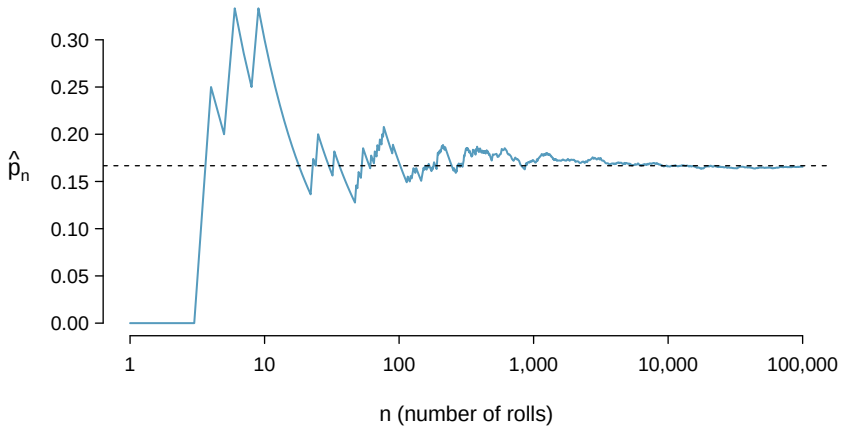


Figure A.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

A.1.1 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**.

Roll a die	$\rightarrow$	`1`, `2`, `3`, `4`, `5`, or `6`
Flip a coin	$\rightarrow$	`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. Let $\hat{p}_n$ be the proportion of outcomes that are `1` after the first n rolls. As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a `1`, $p = 1/6$. Figure A.1 shows this convergence for 100,000 die rolls. The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**.

Law of Large Numbers
As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the 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 A.1. However, these deviations become smaller as the number of rolls increases.

$P(A)$
Probability of outcome A

Above we write p as the probability of rolling a `1`. We can also write this probability as

$$P(\text{rolling a } \texttt{1})$$

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 $P(\text{rolling a } \texttt{1})$ as $P(\texttt{1})$.

⊙ **Guided Practice A.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 A.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.

TIP: Modeling a process as random
It can be helpful to model a process as random even if it is not truly random.

A.1.2 Disjoint or mutually exclusive outcomes

Two outcomes are called **disjoint** or **mutually exclusive** if they cannot both happen. For instance, if we roll a die, the outcomes `1` and `2` are disjoint since they cannot both occur. 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:

$$P(\texttt{1 or 2}) = P(\texttt{1}) + P(\texttt{2}) = 1/6 + 1/6 = 1/3$$

What about the probability of rolling a `1`, `2`, `3`, `4`, `5`, or `6`? Here again, all of the outcomes are disjoint so we add the probabilities:

$$\begin{aligned} &P(\texttt{1 or 2 or 3 or 4 or 5 or 6}) \\ &\quad = P(\texttt{1}) + P(\texttt{2}) + P(\texttt{3}) + P(\texttt{4}) + P(\texttt{5}) + P(\texttt{6}) \\ &\quad = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1. \end{aligned}$$

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

$$P(A_1) + P(A_2) + \cdots + P(A_k) \tag{A.7}$$

[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`.

⊙ **Guided Practice A.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 A.9** In the `email` data set in Chapter 1, 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 $\{\texttt{1}, \texttt{2}\}$ and $B = \{\texttt{4}, \texttt{6}\}$. 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 A.2.

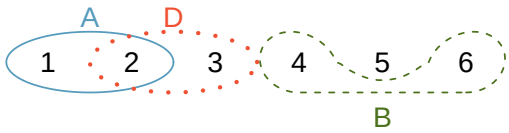


Figure A.2: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

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 A.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 A.11** (a) Using Figure A.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 A.12** In Guided Practice A.11, you confirmed B and D from Figure A.2 are disjoint. Compute the probability that either event B or event D occurs.[6]

---

[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) $P(\texttt{small or big}) = P(\texttt{small}) + P(\texttt{big}) = 0.721 + 0.139 = 0.860$.

[4] (a) $P(A) = P(\texttt{1 or 2}) = P(\texttt{1}) + P(\texttt{2}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (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: $P(B \text{ or } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

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♠

Table A.3: Representations of the 52 unique cards in a deck.

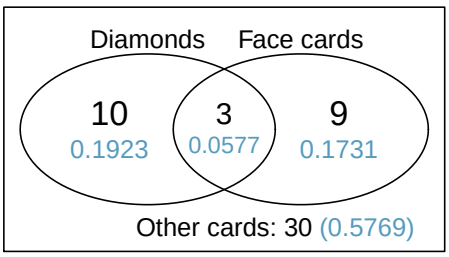


Figure A.4: A Venn diagram for diamonds and face cards.

A.1.3 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 Table A.3. If you are unfamiliar with the cards in a regular deck, please see the footnote.[7]

⊙ **Guided Practice A.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 A.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. $10/52 = 0.1923$).

⊙ **Guided Practice A.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

[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 $P(\diamondsuit) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

[9] The Venn diagram shows face cards split up into "face card but not ♢" and "face card and ♢". Since these correspond to disjoint events, $P(\text{face card})$ is found by adding the two corresponding probabilities: $\frac{3}{52} + \frac{9}{52} = \frac{12}{52} = \frac{3}{13}$.

disjoint events. Instead we use the Venn diagram. We start by adding the probabilities of the two events:

$$P(A) + P(B) = P(\diamondsuit) + P(\text{face card}) = 12/52 + 13/52$$

However, the three cards that are in both events were counted twice, once in each probability. We must correct this double counting:

$$\begin{aligned} P(A \text{ or } B) &= P(\text{face card or } \diamondsuit) \\ &= P(\text{face card}) + P(\diamondsuit) - P(\text{face card and } \diamondsuit) \\ &= 12/52 + 13/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned} \tag{A.15}$$

Equation (A.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 at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{A.16}$$

where $P(A \text{ and } B)$ is the probability that both events occur.

TIP: "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.

⊙ **Guided Practice A.17** (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 A.18** 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]

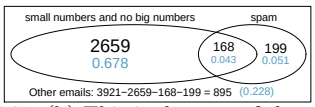
⊙ **Guided Practice A.19** (a) Use your Venn diagram from Guided Practice A.18 to determine the probability a randomly drawn email from the `email` data set is spam and had small numbers (but not big numbers). (b) What is the probability that the email had either of these attributes?[12]

A.1.4 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Table A.5 shows the probability distribution for the sum of two dice.

[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 (A.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: $0.678 + 0.043 + 0.051 = 0.772$.

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table A.5: 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

Table A.6: Proposed distributions of US household incomes (Guided Practice A.20).

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 A.20** Table A.6 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?[13]

Chapter 1 emphasized the importance of plotting data to provide quick summaries. Probability distributions can also be summarized in a bar plot. For instance, the distribution of US household incomes is shown in Figure A.7 as a bar plot.[14] The probability distribution for the sum of two dice is shown in Table A.5 and plotted in Figure A.8.

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 bar plot that resembles 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 A.19 on page 316.

A.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.

S
Sample space

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D 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 A.9 shows the relationship between D, D^c, and the sample space S.

A^c
Complement of outcome A

[13] 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).

[14] It is also possible to construct a distribution plot when income is not artificially binned into four groups.

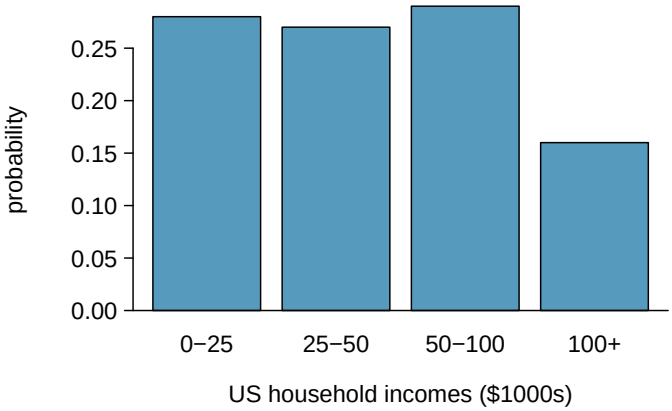


Figure A.7: The probability distribution of US household income.

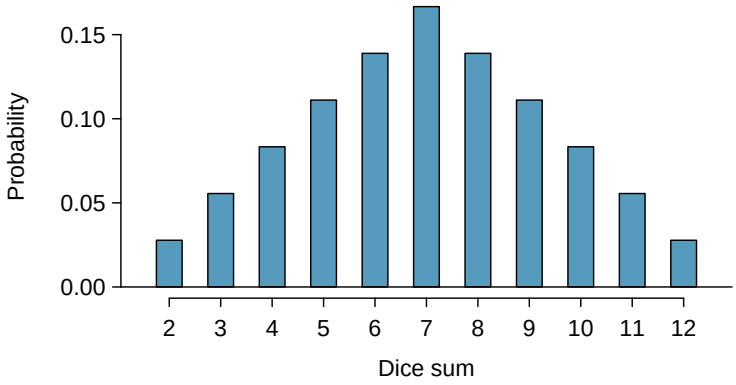


Figure A.8: The probability distribution of the sum of two dice.

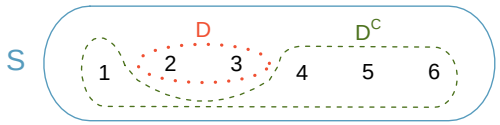


Figure A.9: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

Guided Practice A.21 (a) Compute $P(D^c) = P(\text{rolling a } 1, 4, 5, \text{ or } 6)$. (b) What is $P(D) + P(D^c)$?[15]

Guided Practice A.22 Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure A.2 on page 298. (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 $P(B) + P(B^c)$.[16]

A complement of an event A is constructed to have two very important properties: (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. Property (i) implies

$$P(A \text{ or } A^c) = 1 \tag{A.23}$$

That is, if the outcome is not in A, it must be represented in A^c. We use the Addition Rule for disjoint events to apply Property (ii):

$$P(A \text{ or } A^c) = P(A) + P(A^c) \tag{A.24}$$

Combining Equations (A.23) and (A.24) yields a very useful relationship between the probability of an event and its complement.

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:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c) \tag{A.25}$$

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 A.26 Let A represent the event where we roll two dice and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Table A.5 on page 301. (c) Determine $P(A)$.[17]

Guided Practice A.27 Consider again the probabilities from Table A.5 and rolling two dice. Find the following probabilities: (a) The sum of the dice is *not* 6. (b) The sum is at least 4. That is, determine the probability of the event $B = \{4, 5, ..., 12\}$. (c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, ..., 10\}$.[18]

[15] (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, $P(D) + P(D^c) = 1$.

[16] 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 $P(B) + P(B^c) = 1$.

[17] (a) The complement of A: when the total is equal to 12. (b) $P(A^c) = 1/36$. (c) Use the probability of the complement from part (b), $P(A^c) = 1/36$, and Equation (A.25): $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

[18] (a) First find $P(6) = 5/36$, then use the complement: $P(\text{not } 6) = 1 - P(6) = 31/36$.

A.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 A.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 A.5 (page 295), 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 A.10. 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.

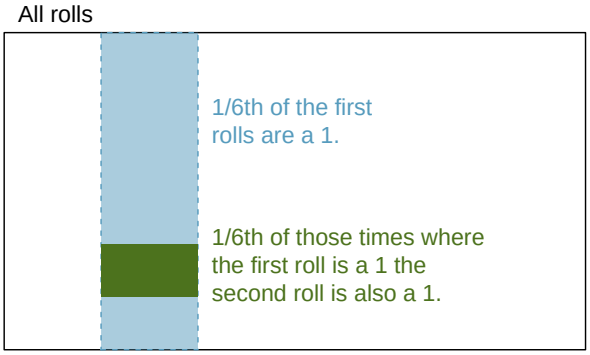


Figure A.10: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

Example A.28 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 A.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:

$$\begin{aligned} P(white = 1 \text{ and } red = 1 \text{ and } blue = 1) &= P(white = 1) \times P(red = 1) \times P(blue = 1) \\ &= (1/6) \times (1/6) \times (1/6) = 1/216 \end{aligned}$$

Examples A.5 and A.28 illustrate what is called the Multiplication Rule for independent processes.

(b) First find the complement, which requires much less effort: $P(2 \text{ or } 3) = 1/36 + 2/36 = 1/12$. Then calculate $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$.

(c) As before, finding the complement is the clever way to determine $P(D)$. First find $P(D^c) = P(11 \text{ or } 12) = 2/36 + 1/36 = 1/12$. Then calculate $P(D) = 1 - P(D^c) = 11/12$.

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:

$$P(A \text{ and } B) = P(A) \times P(B) \tag{A.29}$$

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 A.30** 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?[19]

⊙ **Guided Practice A.31** Suppose 5 people are selected at random.[20]

(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?

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[21] using the Multiplication Rule:

$$\begin{aligned} P(\text{right-handed and female}) &= P(\text{right-handed}) \times P(\text{female}) \\ &= 0.91 \times 0.50 = 0.455 \end{aligned}$$

[19](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 $0.91 \times 0.91 = 0.8281$.

[20](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:

$$\begin{aligned} P(\text{all five are RH}) &= P(\text{first} = \texttt{RH}, \text{second} = \texttt{RH}, \ldots, \text{fifth} = \texttt{RH}) \\ &= P(\text{first} = \texttt{RH}) \times P(\text{second} = \texttt{RH}) \times \cdots \times P(\text{fifth} = \texttt{RH}) \\ &= 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624 \end{aligned}$$

(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:

$$P(\text{not all RH}) = 1 - P(\text{all RH}) = 1 - 0.624 = 0.376$$

[21]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.

⊙ **Guided Practice A.32** Three people are selected at random.[22]

(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 (A.29).

● **Example A.33** 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 A.29 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.

A.2 Conditional probability

Are students more likely to use marijuana when their parents used drugs? The `drug_use` data set contains a sample of 445 cases with two variables, `student` and `parents`, and is summarized in Table A.11.[23] The `student` variable is either `uses` or `not`, where a student is labeled as `uses` if she has recently used marijuana. The `parents` variable takes the value `used` if at least one of the parents used drugs, including alcohol.

		parents		
		used	not	Total
student	uses	125	94	219
	not	85	141	226
	Total	210	235	445

Table A.11: Contingency table summarizing the `drug_use` data set.

● **Example A.34** If at least one parent used drugs, what is the chance their child (`student`) uses?

We will estimate this probability using the data. Of the 210 cases in this data set where `parents` = `used`, 125 represent cases where `student` = `uses`:

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60$$

[22] 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.

[23] Ellis GJ and Stone LH. 1979. Marijuana Use in College: An Evaluation of a Modeling Explanation. Youth and Society 10:323-334.

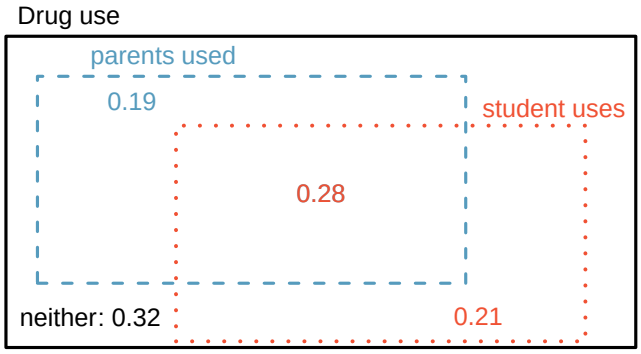


Figure A.12: A Venn diagram using boxes for the `drug_use` data set.

	parents: used	parents: not	Total
student: uses	0.28	0.21	0.49
student: not	0.19	0.32	0.51
Total	0.47	0.53	1.00

Table A.13: Probability table summarizing parental and student drug use.

Example A.35 A student is randomly selected from the study and she does not use drugs. What is the probability that at least one of her parents used?

If the student does not use drugs, then she is one of the 226 students in the second row. Of these 226 students, 85 had at least one parent who used drugs:

$$P(\texttt{parents} = \texttt{used} \text{ given } \texttt{student} = \texttt{not}) = \frac{85}{226} = 0.376$$

A.2.1 Marginal and joint probabilities

Table A.13 includes row and column totals for each variable separately in the `drug_use` data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without conditioning on any other variables. For instance, a probability based solely on the `student` variable is a marginal probability:

$$P(\texttt{student} = \texttt{uses}) = \frac{219}{445} = 0.492$$

A probability of outcomes for two or more variables or processes is called a **joint probability**:

$$P(\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{not}) = \frac{94}{445} = 0.21$$

It is common to substitute a comma for "and" in a joint probability, although either is acceptable.

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 `drug_use` sample. These proportions are computed by dividing each count in Table A.11 by 445 to obtain the proportions in Table A.13. The joint probability distribution of the `parents` and `student` variables is shown in Table A.14.

Joint outcome	Probability
`parents` = `used`, `student` = `uses`	0.28
`parents` = `used`, `student` = `not`	0.19
`parents` = `not`, `student` = `uses`	0.21
`parents` = `not`, `student` = `not`	0.32
Total	1.00

Table A.14: A joint probability distribution for the `drug_use` data set.

⊙ **Guided Practice A.36** Verify Table A.14 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[24]

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability a random student from the study uses drugs is found by summing the outcomes from Table A.14 where `student` = `uses`:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses}) \\ &\quad = P(\texttt{parents} = \texttt{used}, \texttt{student} = \texttt{uses}) + \\ &\qquad\quad P(\texttt{parents} = \texttt{not}, \texttt{student} = \texttt{uses}) \\ &\quad = 0.28 + 0.21 = 0.49 \end{aligned}$$

A.2.2 Defining conditional probability

There is some connection between drug use of parents and of the student: drug use of one is associated with drug use of the other.[25] In this section, we discuss how to use information about associations between two variables to improve probability estimation.

The probability that a random student from the study uses drugs is 0.49. Could we update this probability if we knew that this student's parents used drugs? Absolutely. To do so, we limit our view to only those 210 cases where parents used drugs and look at the fraction where the student uses drugs:

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60$$

We call this a **conditional probability** because we computed the probability under a condition: `parents` = `used`. 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 separate the text inside our probability notation into the outcome of interest and the condition:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) \\ &= P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60 \end{aligned} \tag{A.37}$$

$P(A|B)$
Probability of outcome A given B

The vertical bar "|" is read as *given*.

In Equation (A.37), we computed the probability a student uses based on the condition that

[24] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.28 + 0.19 + 0.21 + 0.32 = 1.00$.

[25] This is an observational study and no causal conclusions may be reached.

at least one parent used as a fraction:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) \\ &\quad = \frac{\#\text{ times } \texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used}}{\#\text{ times } \texttt{parents} = \texttt{used}} \\ &\quad = \frac{125}{210} = 0.60 \end{aligned} \tag{A.38}$$

We considered only those cases that met the condition, `parents = used`, and then we computed the ratio of those cases that satisfied our outcome of interest, the student uses.

Counts are not always available for data, and instead only marginal and joint probabilities may be provided. 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 Equation (A.38) as an example demonstrating this technique.

We considered only those cases that satisfied the condition, `parents = used`. Of these cases, the conditional probability was the fraction who represented the outcome of interest, `student = uses`. Suppose we were provided only the information in Table A.13 on page 307, i.e. only probability data. Then if we took a sample of 1000 people, we would anticipate about 47% or $0.47 \times 1000 = 470$ would meet our information criterion. Similarly, we would expect about 28% or $0.28 \times 1000 = 280$ to meet both the information criterion and represent our outcome of interest. Thus, the conditional probability could be computed:

$$\begin{aligned} P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) &= \frac{\#\ (\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used})}{\#\ (\texttt{parents} = \texttt{used})} \\ &= \frac{280}{470} = \frac{0.28}{0.47} = 0.60 \end{aligned} \tag{A.39}$$

In Equation (A.39), we examine exactly the fraction of two probabilities, 0.28 and 0.47, which we can write as

$$P(\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used}) \quad \text{and} \quad P(\texttt{parents} = \texttt{used}).$$

The fraction of these probabilities represents our general formula for conditional probability.

Conditional Probability

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{A.40}$$

⊙ **Guided Practice A.41** (a) Write out the following statement in conditional probability notation: "*The probability a random case has* `parents = not` *if it is known that* `student = not`". Notice that the condition is now based on the student, not the parent. (b) Determine the probability from part (a). Table A.13 on page 307 may be helpful.[26]

[26] (a) $P(\texttt{parent} = \texttt{not} | \texttt{student} = \texttt{not})$. (b) Equation (A.40) for conditional probability indicates we should first find $P(\texttt{parents} = \texttt{not} \text{ and } \texttt{student} = \texttt{not}) = 0.32$ and $P(\texttt{student} = \texttt{not}) = 0.51$. Then the ratio represents the conditional probability: $0.32/0.51 = 0.63$.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Table A.15: 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

Table A.16: Table proportions for the smallpox data, computed by dividing each count by the table total, 6224.

⊙ **Guided Practice A.42** (a) Determine the probability that one of the parents had used drugs if it is known the student does not use drugs. (b) Using the answers from part (a) and Guided Practice A.41(b), compute

$$P(\texttt{parents} = \texttt{used}|\texttt{student} = \texttt{not}) + P(\texttt{parents} = \texttt{not}|\texttt{student} = \texttt{not})$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[27]

⊙ **Guided Practice A.43** The data indicate that drug use of parents and children are associated. Does this mean the drug use of parents causes the drug use of the students?[28]

A.2.3 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.[29] 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 A.15 and A.16.

⊙ **Guided Practice A.44** Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[30]

[27] (a) This probability is $\frac{P(\texttt{parents} = \texttt{used and student} = \texttt{not})}{P(\texttt{student} = \texttt{not})} = \frac{0.19}{0.51} = 0.37$. (b) The total equals 1. (c) Under the condition the student does not use drugs, the parents must either use drugs or not. The complement still appears to work *when conditioning on the same information.*

[28] No. This was an observational study. Two potential confounding variables include income and region. Can you think of others?

[29] Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6)*. Geneva: World Health Organization. ISBN 92-4-156110-6.

[30] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{no}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{no})}{P(\texttt{inoculated} = \texttt{no})} = \frac{0.1356}{0.9608} = 0.1411$.

⊙ **Guided Practice A.45** Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice A.44?[31]

⊙ **Guided Practice A.46** 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?[32]

A.2.4 General multiplication rule

Section A.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

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

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 in Equation (A.40) on page 309.

● **Example A.47** 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 Table A.16. We want to determine

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no})$$

and we are given that

$$P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{no}) = 0.8588$$
$$P(\texttt{inoculated} = \texttt{no}) = 0.9608$$

Among the 96.08% of people who were not inoculated, 85.88% survived:

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Table A.16 at the intersection of `no` and `lived` (with a small rounding error).

⊙ **Guided Practice A.48** Use $P(\texttt{inoculated} = \texttt{yes}) = 0.0392$ and $P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{yes}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[33]

[31] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{yes}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{yes})}{P(\texttt{inoculated} = \texttt{yes})} = \frac{0.0010}{0.0392} = 0.0255$. 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.

[32] Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care. There are other valid answers for part (c).

[33] The answer is 0.0382, which can be verified using Table A.16.

⊙ **Guided Practice A.49** If 97.45% of the people who were inoculated lived, what proportion of inoculated people must have died?[34]

Sum of conditional probabilities

Let A_1, ..., A_k represent all the disjoint outcomes for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$P(A_1|B) + \cdots + P(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$P(A|B) = 1 - P(A^c|B)$$

⊙ **Guided Practice A.50** Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[35]

A.2.5 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 A.51** 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.[36]

We can show in Guided Practice A.51(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 A.52** 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?[37]

[34] There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.45% = 2.55% of the people who were inoculated died.

[35] 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 A.46, 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.)

[36] 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.

[37] 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**.

A.2.6 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 A.17. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

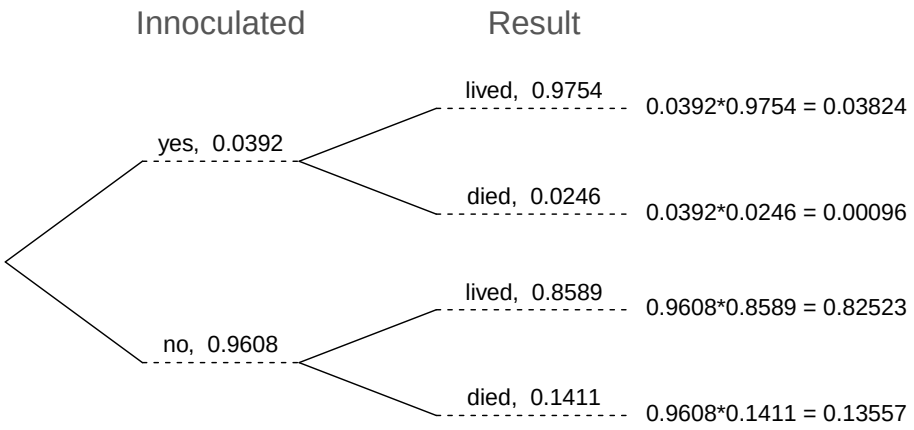


Figure A.17: A tree diagram of the `smallpox` data set.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure A.17. 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 A.17 is the probability that `result` = `lived` conditioned on the information that `inoculated` = `yes`. 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:

$$\begin{aligned} &P(\texttt{inoculated} = \texttt{yes} \text{ and } \texttt{result} = \texttt{lived}) \\ &\quad = P(\texttt{inoculated} = \texttt{yes}) \times P(\texttt{result} = \texttt{lived} | \texttt{inoculated} = \texttt{yes}) \\ &\quad = 0.0392 \times 0.9754 = 0.0382 \end{aligned}$$

Example A.53 Consider the midterm and final for a statistics class. Suppose 13% of students earned an `A` on the midterm. Of those students who earned an `A` on the midterm, 47% received an `A` on the final, and 11% of the students who earned lower than an `A` on the midterm received an `A` on the final. You randomly pick up a final exam and notice the student received an `A`. What is the probability that this student earned an `A` on the midterm?

The end-goal is to find $P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A})$. To calculate this conditional probability, we need the following probabilities:

$$P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) \qquad \text{and} \qquad P(\texttt{final} = \texttt{A})$$

However, this information is not provided, and it is not obvious how to calculate these probabilities. Since we aren't sure how to proceed, it is useful to organize the information into a tree diagram, as shown in Figure A.18. When constructing a tree diagram, variables provided with marginal probabilities are often used to create the tree's primary branches; in this case, the marginal probabilities are provided for midterm grades. The final grades, which correspond to the conditional probabilities provided, will be shown on the secondary branches.

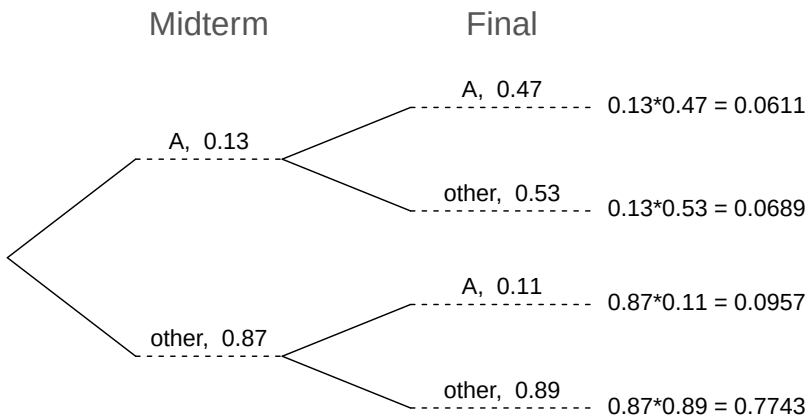


Figure A.18: A tree diagram describing the `midterm` and `final` variables.

With the tree diagram constructed, we may compute the required probabilities:

$$\begin{aligned}
&P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) = 0.0611 \\
&P(\underline{\texttt{final} = \texttt{A}}) \\
&\quad = P(\texttt{midterm} = \texttt{other} \text{ and } \underline{\texttt{final} = \texttt{A}}) + P(\texttt{midterm} = \texttt{A} \text{ and } \underline{\texttt{final} = \texttt{A}}) \\
&\quad = 0.0611 + 0.0957 = 0.1568
\end{aligned}$$

The marginal probability, $P(\texttt{final} = \texttt{A})$, was calculated by adding up all the joint probabilities on the right side of the tree that correspond to $\texttt{final} = \texttt{A}$. We may now finally take the ratio of the two probabilities:

$$\begin{aligned}
P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A}) &= \frac{P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A})}{P(\texttt{final} = \texttt{A})} \\
&= \frac{0.0611}{0.1568} = 0.3897
\end{aligned}$$

The probability the student also earned an A on the midterm is about 0.39.

⊙ **Guided Practice A.54** 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 randomly selected student passed? (c) Compute the probability a student is able to construct a tree diagram if it is known that she passed.[38]

A.3 Random variables

● **Example A.55** 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 A.56** Would you be surprised if the bookstore sold slightly more or less than 105 books?[39]

● **Example A.57** 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

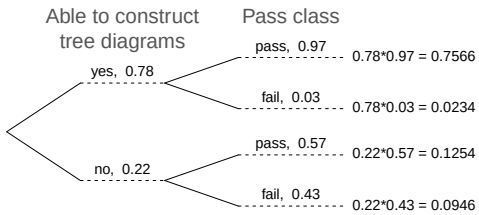
$$\$137 \times 55 = \$7,535$$

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.

[38] (a) The tree diagram is shown to the right. (b) Identify which two joint probabilities represent students who passed, and add them: $P(\text{passed}) = 0.7566 + 0.1254 = 0.8820$. (c) $P(\text{construct tree diagram} \mid \text{passed}) = \frac{0.7566}{0.8820} = 0.8578$.



[39] If they sell a little more or a little less, this should not be a surprise. Hopefully Chapter 1 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.

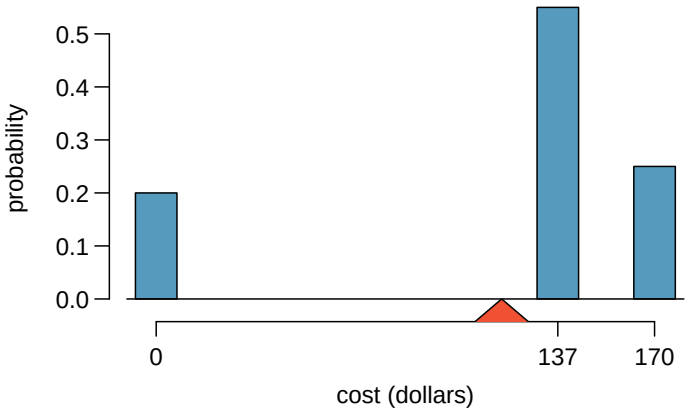


Figure A.19: Probability distribution for the bookstore's revenue from a single student. The distribution balances on a triangle representing the average revenue per student.

i	1	2	3	Total
x_i	\$0	\$137	\$170	–
$P(X = x_i)$	0.20	0.55	0.25	1.00

Table A.20: The probability distribution for the random variable X, representing the bookstore's revenue from a single student.

Example A.58 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 $\$11,785/100 = \$117.85$.

A.3.1 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 A.19 and Table A.20.

$E(X)$
Expected value of X

We computed the average outcome of X as \$117.85 in Example A.58. 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:

$$E(X) = 0 \times P(X = 0) + 137 \times P(X = 137) + 170 \times P(X = 170)$$
$$= 0 \times 0.20 + 137 \times 0.55 + 170 \times 0.25 = 117.85$$

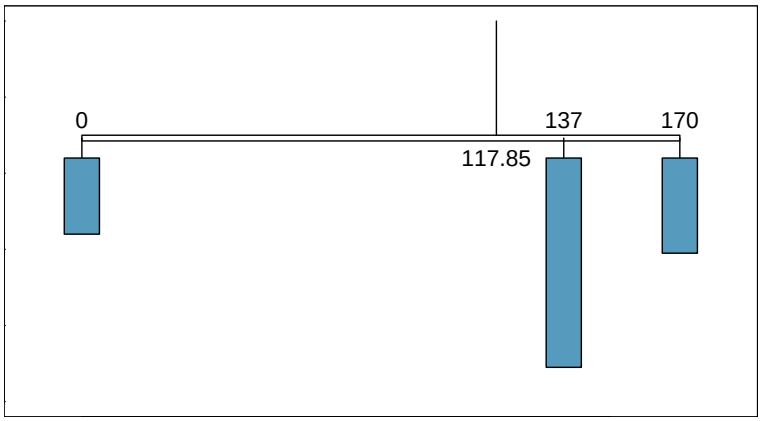


Figure A.21: A weight system representing the probability distribution for X. The string holds the distribution at the mean to keep the system balanced.

Expected value of a Discrete Random Variable

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + \cdots + x_k \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \end{aligned} \tag{A.59}$$

The Greek letter μ may be used in place of the notation $E(X)$.

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 $\mu = 117.85$.

It is also possible to compute the expected value of a continuous random variable. However, it requires a little calculus and we save it for a later class.[40]

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 A.19 and A.21. The idea of a center of gravity also expands to continuous probability distributions. Figure A.22 shows a continuous probability distribution balanced atop a wedge placed at the mean.

A.3.2 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 1.6.4 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

[40] $\mu = \int x f(x) dx$ where $f(x)$ represents a function for the density curve.

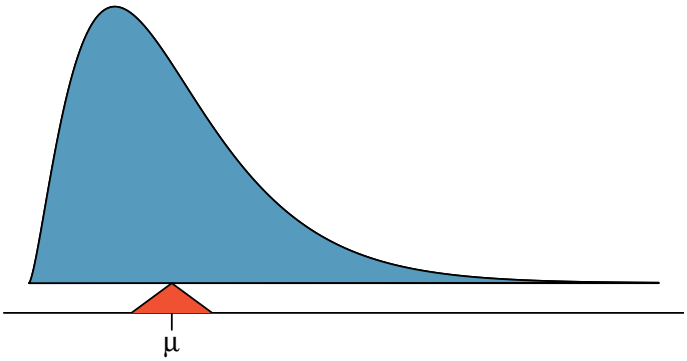


Figure A.22: A continuous distribution can also be balanced at its mean.

we calculate the standard deviation by taking the square root of the variance, just as we did in Section 1.6.4.

$Var(X)$
Variance of X

General variance formula

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by $Var(X)$ or the symbol σ^2, is

$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 \times P(X = x_1) + \cdots \\ &\qquad \cdots + (x_k - \mu)^2 \times P(X = x_k) \\ &= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j) \end{aligned} \tag{A.60}$$

The standard deviation of X, labeled σ, is the square root of the variance.

Example A.61 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	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \times P(X = x_i)$	0	75.35	42.50	117.85

Thus, the expected value is $\mu = 117.85$, which we computed earlier. The variance can be constructed by extending this table:

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \times P(X = x_i)$	0	75.35	42.50	117.85
$x_i - \mu$	-117.85	19.15	52.15	
$(x_i - \mu)^2$	13888.62	366.72	2719.62	
$(x_i - \mu)^2 \times P(X = x_i)$	2777.7	201.7	679.9	3659.3

The variance of X is $\sigma^2 = 3659.3$, which means the standard deviation is $\sigma = \sqrt{3659.3} = \$60.49$.

⊙ **Guided Practice A.62** 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.[41]

(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.

A.3.3 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 A.63** 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.

[41] (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 $y_i \times P(Y = y_i)$. The result of part (d) is the square-root of the variance listed on in the total on the last line: $\sigma = \sqrt{Var(Y)} = \$69.28$.

i (scenario)	1 (`noBook`)	2 (`textbook`)	3 (`both`)	Total
y_i	0.00	159.00	200.00	
$P(Y = y_i)$	0.15	0.25	0.60	
$y_i \times P(Y = y_i)$	0.00	39.75	120.00	$E(Y) = 159.75$
$y_i - E(Y)$	-159.75	-0.75	40.25	
$(y_i - E(Y))^2$	25520.06	0.56	1620.06	
$(y_i - E(Y))^2 \times P(Y)$	3828.0	0.1	972.0	$Var(Y) \approx 4800$

Example A.64 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 A.65** 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.[42]

⊙ **Guided Practice A.66** 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?[43]

⊙ **Guided Practice A.67** Would you be surprised if John's weekly commute wasn't exactly 90 minutes or if Elena didn't make exactly \$152? Explain.[44]

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.[45]

[42] She will make X dollars on the TV but spend Y dollars on the toaster oven: $X - Y$.

[43] $E(X - Y) = E(X) - E(Y) = 175 - 23 = \$152$. She should expect to make about \$152.

[44] 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.

[45] 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.

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

$$aX + bY \tag{A.68}$$

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:

$$a \times E(X) + b \times E(Y)$$

Recall that the expected value is the same as the mean, e.g. $E(X) = \mu_X$.

Example A.69 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 A.70 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.[46]

Guided Practice A.71 You should have found that Leonard expects a positive gain in Guided Practice A.70. However, would you be surprised if he actually had a loss this month?[47]

A.3.4 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 A.70. 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 A.23 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.

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 variances of each stock's monthly return will be useful, and these are shown in Table A.24. The stocks' returns are nearly independent.

[46] $E(\$6000 \times X + \$2000 \times Y) = \$6000 \times 0.021 + \$2000 \times 0.004 = \$134$.
[47] No. While stocks tend to rise over time, they are often volatile in the short term.

Figure A.23: 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.

	Mean ($\bar{x}$)	Standard deviation (s)	Variance (s^2)
GOOG	0.0210	0.0846	0.0072
XOM	0.0038	0.0519	0.0027

Table A.24: 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.

Here we use an equation from probability theory to describe the uncertainty of Leonard's monthly returns; we leave the proof of this method to a dedicated probability course. The variance of a linear combination of random variables can be computed by plugging in the variances of the individual random variables and squaring the coefficients of the random variables:

$$Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

It is important to note that this equality assumes the random variables are independent; if independence doesn't hold, then more advanced methods are necessary. This equation can be used to compute the variance of Leonard's monthly return:

$$\begin{aligned} Var(6000 \times X + 2000 \times Y) &= 6000^2 \times Var(X) + 2000^2 \times Var(Y) \\ &= 36,000,000 \times 0.0072 + 4,000,000 \times 0.0027 \\ &= 270,000 \end{aligned}$$

The standard deviation is computed as the square root of the variance: $\sqrt{270,000} = \$520$. 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.

Variability of linear combinations of random variables

The variance of a linear combination of random variables may be computed by squaring the constants, substituting in the variances for the random variables, and computing the result:

$$Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

This equation is valid as long as the random variables are independent of each other. The standard deviation of the linear combination may be found by taking the square root of the variance.

Example A.72 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 was

$$X_1 + X_2 + X_3 + X_4 + X_5$$

Each coefficient is 1, and the variance of each day's time is $4^2 = 16$. Thus, the variance of the total weekly commute time is

$$\text{variance} = 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 = 5 \times 16 = 80$$

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{80} = 8.94$$

The standard deviation for John's weekly work commute time is about 9 minutes.

Guided Practice A.73 The computation in Example A.72 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.[48]

Guided Practice A.74 Consider Elena's two auctions from Guided Practice A.65 on page 320. 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.[49]

Consider again Guided Practice A.74. 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.

[48] 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.

[49] The equation for Elena can be written as

$$(1) \times X + (-1) \times Y$$

The variances of X and Y are 625 and 64. We square the coefficients and plug in the variances:

$$(1)^2 \times Var(X) + (-1)^2 \times Var(Y) = 1 \times 625 + 1 \times 64 = 689$$

The variance of the linear combination is 689, and the standard deviation is the square root of 689: about \$26.25.

Appendix B

End of chapter exercise solutions

1 Introduction to data

1.1 (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. Control: $2/46 = 0.04 \rightarrow 4\%$. (b) There is a 19% difference between the pain reduction rates in the two groups. At first glance, it appears patients in the treatment group are more likely to experience pain reduction from the acupuncture treatment. (c) Answers may vary but should be sensible. Two possible answers: [1]Though the groups' difference is big, I'm skeptical the results show a real difference and think this might be due to chance. [2]The difference in these rates looks pretty big, so I suspect acupuncture is having a positive impact on pain.

1.3 (a-i) 143,196 eligible study subjects born in Southern California between 1989 and 1993. (a-ii) 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. These are continuous numerical variables. (a-iii) The research question: "Is there an association between air pollution exposure and preterm births?" (b-i) 600 adult patients aged 18-69 years diagnosed and currently treated for asthma. (b-ii) The variables were whether or not the patient practiced the Buteyko method (categorical) and measures of quality of life, activity, asthma symptoms and medication reduction of the patients (categorical, ordinal). It may also be reasonable to treat the ratings on a scale of 1 to 10 as discrete numerical variables. (b-iii) The research question: "Do asthmatic patients who practice the Buteyko method experience improvement in their condition?"

1.5 (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*.

1.7 (a) Population of interest: all births in Southern California. Sample: 143,196 births between 1989 and 1993 in Southern California. If births in this time span 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 do not imply causal relationships. (b) Population: all 18-69 year olds diagnosed and currently treated for asthma. Sample: 600 adult patients aged 18-69 years diagnosed and currently treated for asthma. Since the sample consists of voluntary patients, the results cannot necessarily be generalized to the population at large. However, since the study is an experiment, the findings can be used to establish causal relationships.

1.9 (a) Explanatory: number of study hours per week. Response: GPA. (b) There is a slight positive relationship between the two variables. One respondent reported a GPA above 4.0, which is a data error. There are also a few respondents who reported unusually high study hours (60 and 70 hours/week). The variability in GPA also appears to be larger for stu-

dents who study less than those who study more. Since the data become sparse as the number of study hours increases, it is somewhat difficult to evaluate the strength of the relationship and also the variability across different numbers of study hours. (c) Observational. (d) Since this is an observational study, a causal relationship is not implied.

1.11 (a) Observational. (b) The professor suspects students in a given section may have similar feelings about the course. To ensure each section is reasonably represented, she may choose to randomly select a fixed number of students, say 10, from each section for a total sample size of 40 students. Since a random sample of fixed size was taken within each section in this scenario, this represents stratified sampling.

1.13 Sampling from the phone book would miss unlisted phone numbers, so this would result in bias. People who do not have their numbers listed may share certain characteristics, e.g. consider that cell phones are not listed in phone books, so a sample from the phone book would not necessarily be a representative of the population.

1.15 The estimate will be biased, and it will tend to overestimate the true family size. For example, suppose we had just two families: the first with 2 parents and 5 children, and the second with 2 parents and 1 child. Then if we draw one of the six children at random, 5 times out of 6 we would sample the larger family

1.17 (a) No, this is an observational study. (b) This statement is not justified; it implies a causal association between sleep disorders and bullying. However, this was an observational study. A better conclusion would be "School children identified as bullies are more likely to suffer from sleep disorders than non-bullies."

1.19 (a) Experiment, as the treatment was assigned to each patient. (b) Response: Duration of the cold. Explanatory: Treatment, with 4 levels: *placebo*, *1g*, *3g*, *3g with additives*. (c) Patients were blinded. (d) Double-blind with respect to the researchers evaluating the patients, but the nurses who briefly interacted with patients during the distribution of the medication were not blinded. We could say the study was partly double-blind. (e) No. The patients were randomly assigned to treatment groups and were blinded, so we would expect about an equal number of patients in each group to not adhere to the treatment.

1.21 (a) Experiment. (b) Treatment is exercise twice a week. Control is no exercise. (c) Yes, the blocking variable is age. (d) No. (e) This is an experiment, so a causal conclusion is reasonable. Since the sample is random, the conclusion can be generalized to the population at large. However, we must consider that a placebo effect is possible. (f) Yes. Randomly sampled people should not be required to participate in a clinical trial, and there are also ethical concerns about the plan to instruct one group not to participate in a healthy behavior, which in this case is exercise.

1.23 (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) Association would still be positive. (c) No, they are not independent. See part (a).

1.25 (a) 1/linear and 3/nonlinear. (b) 4/some curvature (nonlinearity) may be present on the right side. "Linear" would also be acceptable for the type of relationship for plot 4. (c) 2.

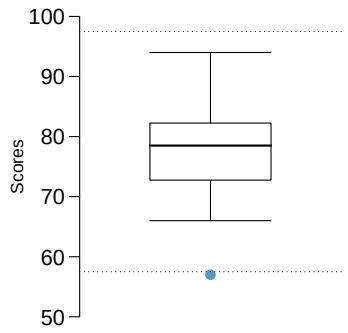
1.27 (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24 + 1) = 73.6$. There are other ways to solve this exercise that do not use a weighted mean. (c) The new score is more than 1 standard deviation away from the previous mean, so increase.

1.29 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.

1.31 (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.

1.33 (a) Both distributions have the same median and IQR. (b) Second distribution has a higher median and higher IQR. (c) Second distribution has higher median. IQRs are equal. (d) Second distribution has higher median and larger IQR.

1.35



1.37 Descriptions will vary a little. (a) 2. Unimodal, symmetric, centered at 60, standard deviation of roughly 3. (b) 3. Symmetric and approximately evenly distributed from 0 to 100. (c) 1. Right skewed, unimodal, centered at about 1.5, with most observations falling between 0 and 3. A very small fraction of observations exceed a value of 5.

1.39 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.

1.41 (a) The median is better; the mean is substantially affected by the two extreme observations. (b) The IQR is better; the standard deviation, like the mean, is substantially affected by the two high salaries.

1.43 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.

1.45 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 0 minutes, which may explain why the average commute time for these counties is relatively low.

1.47 (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.

1.49 The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates the variables are dependent.

2 Foundation for inference

2.1 (a) False. Instead of comparing counts, we should compare percentages. (b) True. (c) False. We cannot infer a causal relationship from an association in an observational study. However, we can say the drug a person is on affects his risk in this case, as he chose that drug and his choice may be associated with other variables, which is why part (b) is true. The difference in these statements is subtle but important. (d) True.

2.3 (a) Proportion who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$ (b) Expected number of cardiovascular problems 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 rate of cardiovascular problems in the study: $67,593 \times \frac{7,979}{227,571} \approx 2370$. (c-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. [Okay if framed a little more generally!] (c-ii) A higher number of patients with cardiovascular problems in the rosiglitazone group than expected under the assumption of independence would provide support for the alternative hypothesis. This would suggest that rosiglitazone increases the risk of such problems. (c-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 all but one or two simulations, which suggests that the actual results did not come from the independence

model. That is, the analysis provides strong evidence that the variables are not independent, and we reject the independence model in favor of the alternative. The study's results provide strong evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

2.5 The subscript $_{pr}$ corresponds to provocative and $_{con}$ to conservative. (a) $H_0 : p_{pr} = p_{con}$. $H_A : p_{pr} \neq p_{con}$. (b) -0.35. (c) The left tail for the p-value is calculated by adding up the two left bins: $0.005 + 0.015 = 0.02$. Doubling the one tail, the p-value is 0.04. (Students may have approximate results, and a small number of students may have a p-value of about 0.05.) Since the p-value is low, we reject H_0. The data provide strong evidence that people react differently under the two scenarios.

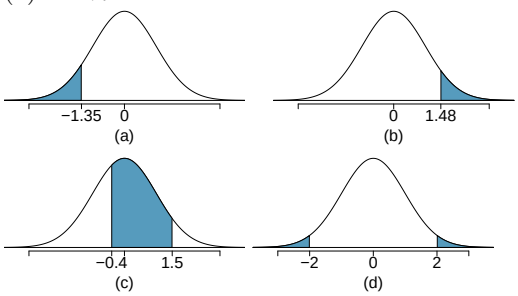
2.7 The primary concern is confirmation bias. If researchers look only for what they suspect to be true using a one-sided test, then they are formally excluding from consideration the possibility that the opposite result is true. Additionally, if other researchers believe the opposite possibility might be true, they would be very skeptical of the one-sided test.

2.9 (a) $H_0 : p = 0.69$. $H_A : p \neq 0.69$. (b) $\hat{p} = \frac{17}{30} = 0.57$. (c) The success-failure condition is not satisfied; note that it is appropriate to use the null value ($p_0 = 0.69$) to compute the expected number of successes and failures. (d) Answers may vary. Each student can be represented with a card. Take 100 cards, 69 black cards representing those who follow the news about Egypt and 31 red cards representing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws) 30 cards representing the 30 high school students. Calculate the proportion of black cards in this sample, $\hat{p}_{sim}$, i.e. the proportion of those who follow the news in the simulation. Repeat this many times (e.g. 10,000 times) and plot the resulting sample proportions. The p-value will be two times the proportion of simulations where $\hat{p}_{sim} \leq 0.57$. (Note: we would generally use a computer to perform these simulations.) (e) The p-value is about $0.001 + 0.005 + 0.020 + 0.035 + 0.075 = 0.136$, meaning the two-sided p-value is about 0.272. Your p-value may vary slightly since it is based on a visual estimate. Since the p-value is greater than 0.05, we fail to reject H_0. The data do not provide strong evidence that the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

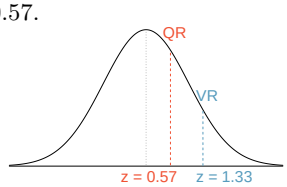
2.11 Each point represents a sample proportion from a simulation. The distributions become (1) smoother / look less discrete, (2) have less variability, and (3) more symmetric (vs. right-skewed). One characteristic that does not change: the distributions are all centered at the same location, $p = 0.1$.

2.13 Each point represents a sample proportion from a simulation. The distributions become (1) smoother / look less discrete, (2) have less variability, and (3) more symmetric (vs. left-skewed). One characteristic that does not change: the distributions are all centered at the same location, $p = 0.95$.

2.15 (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.



2.17 (a) Verbal: $N(\mu = 462, \sigma = 119)$, Quant: $N(\mu = 584, \sigma = 151)$. (b) $Z_{VR} = 1.33$, $Z_{QR} = 0.57$.



(c) She scored 1.33 standard deviations above the mean on the Verbal Reasoning section and 0.57 standard deviations above the mean on the Quantitative Reasoning section. (d) She did better on the Verbal Reasoning section since her Z score on that section was higher. (e) $Perc_{VR} = 0.9082 \approx 91\%$, $Perc_{QR} = 0.7157 \approx 72\%$. (f) $100\% - 91\% = 9\%$ did better than her on VR, and $100\% - 72\% = 28\%$ did better than her on QR. (g) 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. (h) 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)-(f) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

2.19 (a) $Z = 0.84$, which corresponds to 711 on QR. (b) $Z = -0.52$, which corresponds to 400 on VR.

2.21 (a) $Z = 1.2 \rightarrow 0.1151$. (b) $Z = -1.28 \rightarrow 70.6^\circ$F or colder.

2.23 (a) $N(25, 2.78)$. (b) $Z = 1.08 \rightarrow 0.1401$. (c) 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.)

2.25 (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \rightarrow \sigma = \$223.88$.

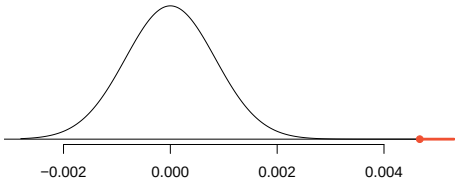
2.27 $Z = 1.56 \rightarrow 0.0594$, i.e. 6%.

2.29 (a) $Z = 0.73 \rightarrow 0.2327$. (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 10^{th} percentile: $Z = -1.28 \rightarrow \$69.80$.

2.31 14/20 = 70% are within 1 SD. Within 2 SD: 19/20 = 95%. Within 3 SD: 20/20 = 100%. They follow this rule closely.

2.33 The distribution is unimodal and symmetric. The superimposed normal curve approximates the distribution pretty well. The points on the normal probability plot also follow a relatively straight line. There is one slightly distant observation on the lower end, but it is not extreme. The data appear to be reasonably approximated by the normal distribution.

2.35 (a) 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. $p_r - p_p = 0$. 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. $p_r - p_p \neq 0$. (b) $\hat{p}_r - \hat{p}_p = \frac{2593}{67593} - \frac{5386}{159978} = 0.00469$. (c) First, draw a picture. Here just the one tail is shown, but the p-value will be both tails:



The Z score is given by $Z = \frac{0.00469-0}{0.00084} = 5.58$. This value is so large that the one tail area is off the normal probability table, so it's area is ≤ 0.0002, so the two-tailed area for the p-value is ≤ 0.0004 (double the one tail area!). (d) We reject the null hypothesis. The data provide strong evidence that there is a difference in the rates of cardiovascular disease for Rosiglitazon and Pioglitazone and that the rate is higher in Rosiglitazon.

2.37 Recall that the general formula is

$$\text{point estimate} \pm z^\star \times SE$$

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:

$$45\% \pm 1.96 \times 1.2\% \quad \rightarrow \quad (42.6\%, 47.4\%)$$

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%.

2.39 (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 2.37, 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.

3 Inference for categorical data

3.1 (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 $1/\sqrt{2}$.

3.3 (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.

3.5 (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 70% $\pm$ 8%. (c) True. By the definition of a confidence interval. (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%.

3.7 With a random sample from $<$ 10% of the population, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times0.44}{600}} = 0.0397 \approx 4\%$

3.9 (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 from less than 10% of the population, 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.

3.11 (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%.

3.13 (a) This is an appropriate setting for a hypothesis test. $H_0 : p = 0.50$. $H_A : p > 0.50$. Both independence and the success-failure condition are satisfied. $Z = 1.12 \rightarrow$ p-value $= 0.1314$. Since the p-value $> \alpha = 0.05$, we fail to reject H_0. The data do not provide strong evidence in favor of the claim. (b) Yes, since we did not reject H_0 in part (a).

3.15 (a) $H_0 : p = 0.38$. $H_A : p \neq 0.38$. Independence (random sample, $<$ 10% of population) 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.

3.17 (a) $H_0 : p = 0.5$. $H_A : p > 0.5$. Independence (random sample, $<$ 10% of population) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ p-value $= 0.0018$. Since the p-value < 0.05, we reject the null hypothesis. The data provide strong evidence that 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 randomly guess, the probability of getting a random sample of 80 people where 53 or more iden-

tify a soda correctly would be 0.0018.

3.19 (a) Independence is satisfied (random sample from $< 10\%$ of the population), 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.

3.21 The margin of error, which is computed as $z^\star SE$, must be smaller than 0.01 for a 90% confidence level. We use $z^\star = 1.65$ for a 90% confidence level, and we can use the point estimate $\hat{p} = 0.52$ in the formula for SE. $1.65\sqrt{0.52(1-0.52)/n} \leq 0.01$. Therefore, the sample size n must be at least 6,796.

3.23 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.

3.25 (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).

3.27 (a) (0.23, 0.33). We are 95% confident that the proportion of Democrats who support the plan is 23% to 33% higher than the proportion of Independents who do. (b) True.

3.29 (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, $< 10\%$ of the population), and the success-failure condition, which we would check using the pooled proportion ($\hat{p} = 235/827 = 0.284$), 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).

3.31 (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, $< 10\%$ of the population), and the success-failure condition, which we would check using the pooled proportion ($\hat{p} = 286/827 = 0.346$), 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 offshore drilling in California.

3.33 Subscript $_C$ means control group. Subscript $_T$ means truck drivers. (a) $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples, $< 10\%$ of the population), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p} = 70/495 = 0.141$). $Z = -1.58 \rightarrow$ p-value $= 0.1164$. Since the p-value is high, 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.

3.35 (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 ($\hat{p} = 36/240 = 0.15$), is satisfied. $Z = 3.04 \rightarrow$ p-value $= 0.0024$. 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 do not appear to be independent.

3.37 (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.

3.39 (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. Degrees of freedom: $df = k - 1 = 3 - 1 = 2$ is more than 1. (d) $X^2 = 2.32$, $df = 2$, p-value > 0.3. (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.

3.41 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.

$$E_{row\ 1,col\ 1} = 151.5 \quad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \quad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \quad E_{row\ 3,col\ 2} = 110.5$$

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. Degrees of freedom: $df = (R-1) \times (C-1) = (3-1) \times (2-1) = 2$, which is greater than 1. $X^2 = 11.47$, $df = 2$ $\rightarrow 0.001 <$ p-value < 0.005. Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is an association between support for off-shore drilling and having a college degree.

3.43 (a) H_0 : There is no relationship between gender and how informed Facebook users are about adjusting their privacy settings. H_A : There is a relationship between gender and how informed Facebook users are about adjusting their privacy settings. (b) The expected counts:

$$E_{row\ 1,col\ 1} = 296.6 \quad E_{row\ 1,col\ 2} = 369.3$$
$$E_{row\ 2,col\ 1} = 54.8 \quad E_{row\ 2,col\ 2} = 68.2$$
$$E_{row\ 3,col\ 1} = 7.6 \quad E_{row\ 3,col\ 2} = 9.4$$

The sample is random, all expected counts are above 5, and $df = (3-1) \times (2-1) = 2 > 1$, so we may proceed with the test.

3.45 It is not appropriate. There are only 9 successes in the sample, so the success-failure condition is not met.

4 Inference for numerical data

4.1 (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 5$, $t^\star_{20} = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

4.3 The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t^\star_{35} = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.

4.5 (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu \neq 8$ (New Yorkers sleep an amount different than 8 hrs per night on average.) (b) Independence: The sample is random and from less than 10% of New Yorkers. The sample is small, so we will use a t distribution. For this size sample, slight skew is acceptable, and the min/max suggest there is not much skew in the data. $T = -1.75$. $df = 25 - 1 = 24$. (c) $0.05 <$ p-value < 0.10. 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 hrs per night or less is between 0.05 and 0.10. (d) Since p-value > 0.05, we do not reject H_0. The data do not provide strong evidence that New Yorkers sleep an amount different than 8 hours per night on average. (e) Yes, as we rejected H_0.

4.7 $t^\star_{19}$ is 1.73 for a one-tail. We want the lower tail, so set -1.73 equal to the T score, then solve for $\bar{x}$: 56.91.

4.9 (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. (b) H_0 : $\mu_{diff} = 0$ (There is no difference in average daily high temperature between January 1, 1968 and January 1, 2008 in the continental US.) H_A : $\mu_{diff} \neq 0$ (Average daily high temperature in January 1, 1968 is different than the average daily high temperature in January, 2008 in the continental US.) (c) Independence: locations are random and represent less than 10% of all possible locations in the US. We are not given the distribution to check the skew. In practice, we would ask to see the data to check this condition, but here we will move forward under the assumption that it is not strongly skewed. (d) $T = 1.60$, $df = 51 - 1 = 50$. $\rightarrow$ p-value between 0.1, 0.2 (two tails!). (e) Since the p-value $> \alpha$ (since not given use 0.05), fail to reject H_0. The data do not provide strong evidence of a temperature change in the continental US on the two dates. (f) Type 2, since we may have incorrectly failed to reject H_0. There may be an increase, but we were unable to detect it. (g) Yes, since we failed to reject H_0, which had a null value of 0.

4.11 (a) (-0.28, 2.48). (b) We are 95% confident that the average daily high on January 1, 2008 in the continental US was 0.24 degrees *lower* to 2.44 degrees *higher* than the average daily high on January 1, 1968. (c) No, since 0 is included in the interval.

4.13 (a) Each of the 36 mothers is related to exactly one of the 36 fathers (and vice-versa), so there is a special correspondence between the mothers and fathers. (b) H_0 : $\mu_{diff} = 0$. H_A : $\mu_{diff} \neq 0$. Independence: random sample from less than 10% of population. The skew of the differences is, at worst, slight. $T = 2.72$, $df = 36 - 1 = 35$ $\rightarrow$ p-value ≈ 0.01. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average IQ scores of mothers and fathers of gifted children are different, and the data indicate that mothers' scores are higher than fathers' scores for the parents of gifted children.

4.15 Independence: Random samples that are less than 10% of the population. In practice, we'd ask for the data to check the skew (which is not provided), but here we will move forward under the assumption that the skew is not extreme (there is some leeway in the skew for such large samples). Use $t^\star_{999} \approx 1.65$. 90% CI: (0.16, 5.84). We are 90% confident that the average score in 2008 was 0.16 to 5.84 points higher than the average score in 2004.

4.17 (a) $H_0 : \mu_{2008} = \mu_{2004}$ $\rightarrow$ $\mu_{2004} - \mu_{2008} = 0$ (Average math score in 2008 is equal to average math score in 2004.) $H_A : \mu_{2008} \neq \mu_{2004}$ $\rightarrow$ $\mu_{2004} - \mu_{2008} \neq 0$ (Average math score in 2008 is different than average math score in 2004.) Conditions necessary for inference were checked in Exercise 4.15. $T = -1.74$, $df = 999$ $\rightarrow$ p-value between 0.05, 0.10. Since the p-value $< \alpha$, reject H_0. The data provide strong evidence that the average math score for 13 year old students has changed between 2004 and 2008. (b) Yes, a Type 1 error is possible. We rejected H_0, but it is possible H_0 is actually true. (c) No, since we rejected H_0 in part (a).

4.19 (a) We are 95% confident that those on the Paleo diet lose 0.891 pounds less to 4.891 pounds more than those in the control group. (b) No. The value representing no difference between the diets, 0, is included in the confidence interval. (c) The change would have shifted the confidence interval by 1 pound, yielding $CI = (0.109, 5.891)$, which does not include 0. Had we observed this result, we would have rejected H_0.

4.21 The independence condition is satisfied. Almost any degree of skew is reasonable with such large samples. Compute the joint SE: $\sqrt{SE_M^2 + SE_W^2} = 0.114$. The 95% CI: (-11.32, -10.88). We are 95% confident that the average body fat percentage in men is 11.32% to 10.88% lower than the average body fat percentage in women.

4.23 No, he should not move forward with the test since the distributions of total personal income are very strongly skewed. When sample sizes are large, we can be a bit lenient with skew. However, such strong skew observed in this exercise would require somewhat large sample sizes, somewhat higher than 30.

4.25 (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. With fewer than 10 observations, we use the t distribution to model the sample mean. The normal probability plot of the differences shows an approximately straight line. There isn't a clear reason why this distribution would be skewed, and since the normal quantile plot looks reasonable, we can mark this condition as reasonably satisfied. (d) $T = 4.94$ for $df = 10 - 1 = 9 \rightarrow$ p-value < 0.01. (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 6^{th} than on Friday the 13^{th}. (We might believe this intersection is representative of all roads, i.e. there is higher traffic on Friday the 6^{th} relative to Friday the 13^{th}. However, we should be cautious of the required assumption for such a generalization.) (f) If the average number of cars passing the intersection actually was the same on Friday the 6^{th} and 13^{th}, 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.

4.27 (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. $0.02 <$ p-value < 0.05. 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 6^{th} and Friday the 13^{th}. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6^{th} relative to Friday the 13^{th}. (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 13^{th} has a higher chance of harm than on any other night.

4.29 (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. $T = 3.02$, $df = min(11, 9) = 9 \rightarrow 0.01 <$ p-value < 0.02. 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, since we rejected H_0. (d) Yes, since p-value > 0.01, we would have failed to reject H_0.

4.31 $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.48$, $df = 11 \rightarrow$ p-value < 0.01. 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 are can be attributed to the diet.

4.33 $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow 0.02 <$ p-value < 0.05. 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.

4.35 Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} > 0$: Treatment is effective in reducing Pd T scores, the average pre-treatment score is higher than the average post-treatment score. Note that the reported values are pre minus post, so we are looking for a positive difference, which would correspond to a reduction in the psychopathic deviant T score. Conditions are checked as follows. Independence: The subjects are randomly assigned to treatments, so the patients in each group are independent. All three sample sizes are smaller than 30, so we use t tests.Distributions of differences are somewhat skewed. The sample sizes are small, so we cannot reliably relax this assumption. (We will proceed, but we would not report the results of this specific analysis, at least for treatment group 1.) For all three groups: $df = 13$. $T_1 = 1.89$ ($0.025 <$ p-value < 0.05), $T_2 = 1.35$ (p-value $= 0.10$), $T_3 = -1.40$ (p-value > 0.10). The only significant test reduction is found in Treatment 1, however, we had earlier noted that this result might not be reliable due to the skew in the distribution. Note that the calculation of the p-value for Treatment 3 was unnecessary:

the sample mean indicated a increase in Pd T scores under this treatment (as opposed to a decrease, which was the result of interest). That is, we could tell without formally completing the hypothesis test that the p-value would be large for this treatment group.

4.37 H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$. H_A: The average weight varies across some (or all) groups. Independence: Chicks are randomly assigned to feed types (presumably kept separate from one another), therefore independence of observations is reasonable. Approx. normal: the distributions of weights within each feed type appear to be fairly symmetric. Constant variance: Based on the side-by-side box plots, the constant variance assumption appears to be reasonable. There are differences in the actual computed standard deviations, but these might be due to chance as these are quite small samples. $F_{5,65} = 15.36$ and the p-value is approximately 0. With such a small p-value, we reject H_0. The data provide convincing evidence that the average weight of chicks varies across some (or all) feed supplement groups.

4.39 (a) H_0: The population mean of MET for each group is equal to the others. H_A: At least one pair of means is different. (b) Independence: We don't have any information on how the data were collected, so we cannot assess independence. To proceed, we must assume the subjects in each group are independent. In practice, we would inquire for more details. Approx. normal: The data are bound below by zero and the standard deviations are larger than the means, indicating very strong strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable. Constant variance: This condition is sufficiently met, as the standard deviations are reasonably consistent across groups. (c) See below, with the last column omitted:

	Df	Sum Sq	Mean Sq	F value
coffee	4	10508	2627	5.2
Residuals	50734	25564819	504	
Total	50738	25575327		

(d) Since p-value is very small, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.

4.41 (a) H_0: Average GPA is the same for all majors. H_A: At least one pair of means are different. (b) Since p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average GPAs across three groups of majors. (c) The total degrees of freedom is $195 + 2 = 197$, so the sample size is $197 + 1 = 198$.

4.43 (a) False. As the number of groups increases, so does the number of comparisons and hence the modified significance level decreases. (b) True. (c) True. (d) False. We need observations to be independent regardless of sample size.

4.45 (a) H_0: Average score difference is the same for all treatments. H_A: At least one pair of means are different. (b) We should check conditions. If we look back to the earlier exercise, we will see that the patients were randomized, so independence is satisfied. There are some minor concerns about skew, especially with the third group, though this may be acceptable. The standard deviations across the groups are reasonably similar. Since the p-value is less than 0.05, reject H_0. The data provide convincing evidence of a difference between the average reduction in score among treatments. (c) We determined that at least two means are different in part (b), so we now conduct $K = 3 \times 2/2 = 3$ pairwise t tests that each use $\alpha = 0.05/3 = 0.0167$ for a significance level. Use the following hypotheses for each pairwise test. H_0: The two means are equal. H_A: The two means are different. The sample sizes are equal and we use the pooled SD, so we can compute $SE = 3.7$ with the pooled $df = 39$. The p-value only for Trmt 1 vs. Trmt 3 may be statistically significant: $0.01 <$ p-value < 0.02. Since we cannot tell, we should use a computer to get the p-value, 0.015, which is statistically significant for the adjusted significance level. That is, we have identified Treatment 1 and Treatment 3 as having different effects. Checking the other two comparisons, the differences are not statistically significant.

4.47 (a) The point estimate is the sample standard deviation, which is in the table: 9.41 cm. (b) We were told we may assume the data are from a simple random sample, which ensures independence. There are 507 data points (individuals), satisfying the ≥ 30 sample size condition. Lastly, the bootstrap distribution appears to be nearly normal. Therefore, the bootstrap approach is reasonable. (c) A 95% confidence interval may be constructed using the 2.5^{th} and 97.5^{th} percentiles, which are 8.88 and 9.91, respectively. That is, we are 95% confident that the true standard deviation of the population's height is between 8.88 and 9.91 centimeters.

5 Introduction to linear regression

5.1 (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.

5.3 (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.

5.5 (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 for part (b).

5.7 (a) $R = -0.7 \rightarrow (4)$. (b) $R = 0.45 \rightarrow (3)$. (c) $R = 0.06 \rightarrow (1)$. (d) $R = 0.92 \rightarrow (2)$.

5.9 (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.

5.11 (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: $R = 0.636$.

5.13 (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.

5.15 In each part, we may write the husband ages as a linear function of the wife ages: (a) $age_H = age_W + 3$; (b) $age_H = age_W - 2$; and (c) $age_H = 2 \times age_W$. Therefore, 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, such as a data set of 5 women with ages 26, 27, 28, 29, and 30 (or some other set of ages). Then, based on the description, say for part (a), we can compute their husbands' ages as 29, 30, 31, 32, and 33. We can plot these points to see they fall on a straight line, and they always will. The same approach can be applied to the other parts as well.

5.17 (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 that it is not evident in the current data but that is important to consider.

5.19 (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.

5.21 The relationship between the variables is somewhat linear. However, there are two apparent outliers. The residuals do not show a random scatter around 0. A simple linear model may not be appropriate for these data, and we should investigate the two outliers.

5.23 (a) $\sqrt{R^2} = 0.849$. Since the trend is negative, R is also negative: $R = -0.849$. (b) $b_0 = 55.34$. $b_1 = -0.537$. (c) For a neighborhood with 0% reduced-fee lunch, we would expect 55.34% of the bike riders to wear helmets. (d) For every additional percentage point of reduced fee lunches in a neighborhood, we would expect 0.537% fewer kids to be wearing helmets. (e) $\hat{y} = 40 \times (-0.537) + 55.34 = 33.86$, $e = 40 - \hat{y} = 6.14$. There are 6.14% more bike riders wearing helmets than predicted by the regression model in this neighborhood.

5.25 (a) The outlier is in the upper-left corner. Since it is horizontally far from the center of the data, it is a point with high leverage. Since the slope of the regression line would be very different if fit without this point, it is also an influential point. (b) The outlier is located in the lower-left corner. It is horizontally far from the rest of the data, so it is a high-leverage point. The line again would look notably different if the fit excluded this point, meaning it the outlier is influential. (c) The outlier is in the upper-middle of the plot. Since it is near the horizontal center of the data, it is not a high-leverage point. This means it also will have little or no influence on the slope of the regression line.

5.27 (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.

5.29 (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_0: The true slope coefficient of height is greater than zero ($\beta_1 > 0$). A two-sided test would also be acceptable for this application. The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so the p-value for the one-sided hypothesis will be even smaller. That is, 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.

5.31 (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 > 0$. A two-sided test would also be acceptable for this application. The p-value, as reported in the table, is incredibly small. Thus, for a one-sided test, the p-value will also be incredibly small, and we reject H_0. The data provide convincing evidence that wives' and husbands' heights are positively correlated. (b) $\widehat{height}_W = 43.5755 + 0.2863 \times height_H$. (c) Slope: For each additional inch

in husband's height, the average wife's height is expected to be an additional 0.2863 inches on average. Intercept: Men who are 0 inches tall are expected to have wives who are, on average, 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.

5.33 (a) 25.75. (b) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. A one-sided test also may be reasonable for this application. $T = 2.23$, $df = 23 \rightarrow$ p-value between 0.02 and 0.05. So we reject H_0. There is an association between gestational age and head circumference. We can also say that the associaation is positive.

6 Multiple and logistic regression

6.1 (a) $\widehat{baby\_weight} = 123.05 - 8.94 \times smoke$ (b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than babies born to non-smoking mothers. Smoker: $123.05 - 8.94 \times 1 = 114.11$ ounces. Non-smoker: $123.05 - 8.94 \times 0 = 123.05$ ounces. (c) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. $T = -8.65$, and the p-value is approximately 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the true slope parameter is different than 0 and that there is an association between birth weight and smoking. Furthermore, having rejected H_0, we can conclude that smoking is associated with lower birth weights.

6.3 (a) $\widehat{baby\_weight} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$. (b) $\beta_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant. β_{age}: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant. (c) Parity might be correlated with one of the other variables in the model, which complicates model estimation. (d) $\widehat{baby\_weight} = 120.58$. $e = 120 - 120.58 = -0.58$. The model over-predicts this baby's birth weight. (e) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

6.5 (a) (-0.32, 0.16). We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model. (b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

6.7 (a) There is not a significant relationship between the age of the mother. We should consider removing this variable from the model. (b) All other variables are statistically significant at the 5% level.

6.9 Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted R^2 for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward-selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted R^2.)

6.11 Nearly normal residuals: The normal probability plot shows a nearly normal distribution of the residuals, however, there are some minor irregularities at the tails. With a data set so large, these would not be a concern.
Constant variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.
Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.
Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.
All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

6.13 (a) There are a few potential outliers, e.g. on the left in the `total_length` variable, but nothing that will be of serious concern in a data set this large. (b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for `sex_male` changed when we removed the `head_length` variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.

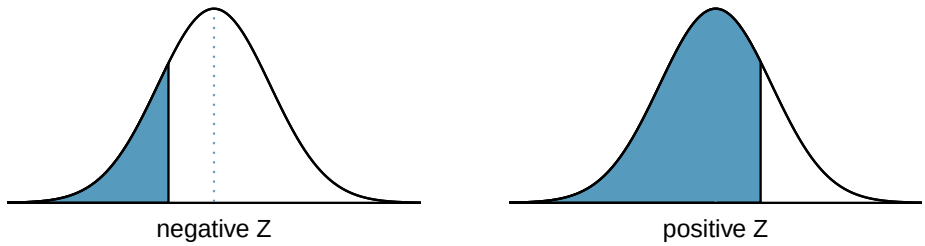
6.15 (a) The logistic model relating $\hat{p}_i$ to the predictors may be written as $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 33.5095 - 1.4207 \times sex\_male_i - 0.2787 \times skull\_width_i + 0.5687 \times total\_length_i$. Only `total_length` has a positive association with a possum being from Victoria. (b) $\hat{p} = 0.0062$. While the probability is very near zero, we have not run diagnostics on the model. We might also be a little skeptical that the model will remain accurate for a possum found in a US zoo. For example, perhaps the zoo selected a possum with specific characteristics but only looked in one region. On the other hand, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)

Appendix C

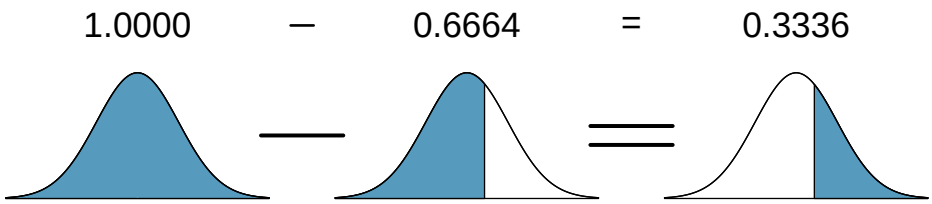
Distribution tables

C.1 Normal Probability Table

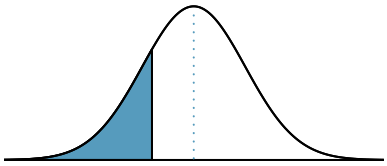
The area to the left of Z represents the percentile of the observation. The normal probability table always lists percentiles.



To find the area to the right, calculate 1 minus the area to the left.

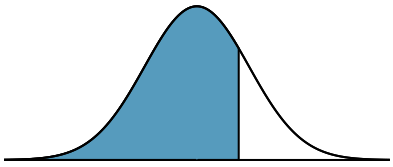


For additional details about working with the normal distribution and the normal probability table, see Section 2.6, which starts on page 85.



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.



	Second decimal place of Z									
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.2 t Distribution Table

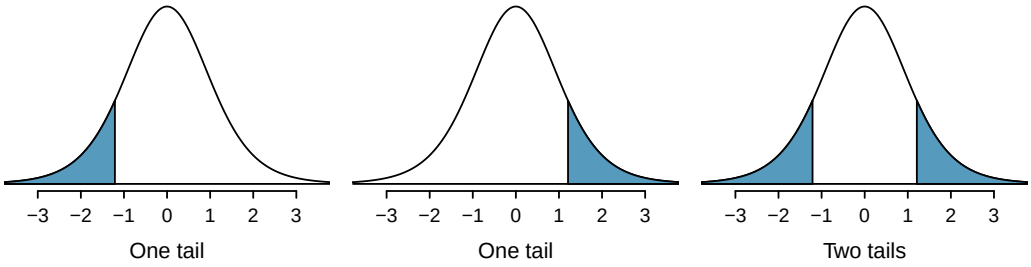


Figure C.1: Three t distributions.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
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

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
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

C.3 Chi-Square Probability Table

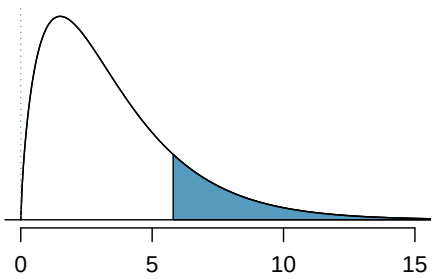


Figure C.2: Areas in the chi-square table always refer to the right tail.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	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