Introductory Statistics for the Life and Biomedical Sciences

Derivative of
OpenIntro Statistics
Third Edition

Original Authors

David M Diez
Christopher D Barr
Mine Çetinkaya-Rundel

Contributing Authors

David Harrington
[Briefly Describe Contribution]

Julie Vu
[Briefly Describe Contribution]

Alice Zhao
[Briefly Describe Contribution]

Contents

Preface

This book provides an introduction to statistics and its applications in the life sciences, and biomedical research. It is based on the freely available *OpenIntro Statistics, Third Edition*, and, like *OpenIntro* it may be downloaded as a free PDF at **Need location**. The text adds substantial new material, revises or eliminates sections from *OpenIntro*, and re-uses some material directly. Readers need not have read *OpenIntro*, since this book is intended to be used independently. We have retained some of the exercises from *OpenIntro* that may not come directly from medicine or the life sciences but illustrate important ideas or methods that are commonly used in fields such as biology.

Introduction to Statistics for the Life and Biomedical Sciences is intended for graduate and undergraduate students interested in careers in biology or medicine, and may also be profitably read by students of public health. It covers many of the traditional introductory topics in statistics used in those fields, but also adds some newer methods being used in molecular biology. Statistics has become an integral part of research in medicine and biology, and the tools for displaying, summarizing and drawing inferences from data are essential both for understanding the outcomes of studies and for incorporating measures of uncertainty into that understanding. An introductory text in statistics for students considering careers in medicine, public health or the life sciences should be more than the usual introduction with more examples from biology or medical science. Along with the value of careful, robust analyses of experimental and observational data, it should convey some of the excitement of discovery that emerges from the interplay of science with data collection and analysis. We hope we have conveyed some of that excitement here.

We have tried to balance the sometimes competing demands of mastering the important technical aspects of methods of analysis with gaining an understanding of important concepts. The examples and exercises include opportunities for students to build skills in conducting data analyses and to state conclusions with clear, direct language that is specific to the context of a problem. We also believe that computing is an essential part of statistics, just as mathematics was when computing was more difficult or expensive. The text includes many examples where software is used to aid in the understanding of the features of a data as well as exercises where computing is used to help illustrate the notions of randomness and variability. Because they are freely available, we use the R statistical language with the *R Studio* interface. Information on downloading R and *R Studio* is may be found in the Labs at **openintro.org**. Nearly all examples and exercises can be adapted to either SAS, Stata or other software, but we have not done that.

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. **Probability (special topic).** The basic principles of probability. An understanding of this chapter is not required for the main content in Chapters **??**-**??**.
3. **Distributions of random variables.** Introduction to the normal model and other key distributions.
4. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population mean.
5. **Inference for numerical data.** Inference for one or two sample means using the normal model and t distribution, and also comparisons of many means using ANOVA.
6. **Inference for categorical data.** Inference for proportions using the normal and chi-square distributions, as well as simulation and randomization techniques.
7. **Introduction to linear regression.** An introduction to regression with two variables. Most of this chapter could be covered after Chapter 1.
8. **Multiple and logistic regression.** An introduction to multiple regression and logistic regression for an accelerated course.

The remainder of this section requires revision

OpenIntro Statistics was written to allow flexibility in choosing and ordering course topics. The material is divided into two pieces: main text and special topics. The main text has been structured to bring statistical inference and modeling closer to the front of a course. Special topics, labeled in the table of contents and in section titles, may be added to a course as they arise naturally in the curriculum.

Examples, exercises, and appendices

Examples and within-chapter 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. Students may find it useful to fill in the bullet after understanding or successfully completing the exercise. Solutions are provided for all within-chapter exercises in footnotes.[1]

There are exercises at the end of each chapter that are useful for practice or homework assignments. Many of these questions have multiple parts, and odd-numbered questions include solutions in Appendix A.

Probability tables for the normal, t, and chi-square distributions are in Appendix B, and PDF copies of these tables are also available from **openintro.org** for anyone to download, print, share, or modify.

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

OpenIntro, online resources, and getting involved

OpenIntro is an organization focused on developing free and affordable education materials. *OpenIntro Statistics*, our first project, is intended for introductory statistics courses at the high school through university levels.

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.[2] 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 particular component 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 volunteers below.

The authors would like to thank Andrew Bray, Meenal Patel, Yongtao Guan, Filipp Brunshteyn, Rob Gould, and Chris Pope for their involvement and contributions. We are also very grateful to Dalene Stangl, Dave Harrington, Jan de Leeuw, Kevin Rader, and Philippe Rigollet for providing us with valuable feedback.

[2] 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? Each of these questions must, of course, be intimately linked to the scientific question at hand.

Many scientific investigations can be conducted with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference. This chapter provides a brief introduction to the basic principles of these areas that will be encountered later in the book, and illustrates the important role statistics plays in medicine and biology.

leaves open the question of why we are using the Counties data

1.1 Case study: using stents to prevent strokes

Section 1.1 introduces an important problem in medicine: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in more detail.

This section describes an experiment (a clinical trial, in the terminology of medical research) designed to assess the effectiveness of stents in treating patients at risk of stroke.[1] Stents are devices placed inside blood vessels and are thought to assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. At the

[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.

time the experiment was designed, it was natural to conjecture that there would be similar benefits for patients at risk of stroke. The study was designed to answer the question:

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

The study team collected data on 451 patients whose clinical condition suggested a high risk for a stroke. Each study volunteer (or study subject, in the terms of medical research) 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 stroke risk factors, and counseling on lifestyle modification.
>
> **Control group**. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point for estimating the effect of stents in the treatment group.

The study team studied the effect of stents at two time points: 30 days and 365 days after randomization. The outcomes for 5 patients are summarized in Table 1.1. Patient outcomes are recorded as “stroke” or “no event”, representing whether or not the patient had a stroke by the end of a time period.

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

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

where are the data for this?

Table 1.2 summarizes subject-level data in a helpful way, showing patterns in the data. For instance, the first column on the left side of the table shows that a total of 46 patients experienced a stroke within 30 days: 33 were in the treatment group, while 13 were in the control group.

	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.

The table makes it possible to compute some simple summary statistics. A **summary statistic** is a single number summarizing a large amount of data.[2] For instance, the

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

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 for examining possible differences between the groups. They show something that surprised and disappointed the study team: an additional 8% of patients in the treatment group experienced a stroke. This illustrates the importance of experimentally verifying scientific hypotheses.

The results also raise an important statistical issue: does the study provide definitive evidence that stents are harmful? In other words, is the 8% difference between the two groups larger than one would expect by chance variation alone?

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 experiment or study. It may well be 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 study size), the less credible it is that the difference is due to chance alone. If out of 100 flips, a coin landed heads up only 5 times, it would be reasonable to doubt that the outcome was due to chance; perhaps the coin is weighted so that tails are more likely to occur.

The material on hypothesis testing will provide the statistical tools to examine this issue. In the stent study, such methods showed compelling evidence that the 8% difference was indeed larger than expected by chance, and that for the types of subjects participating in this clinical trial, stents increased the risk of stroke.

Be careful: the results of this study should not be generalized to all patients and all stents. This study looked at patients with very specific characteristics who may not be representative of all stroke patients. Additionally, there are many types of stents, and this study only considered the self-expanding Wingspan stent (Boston Scientific).

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

The Functional polymorphisms Associated with Muscle Size and Strength study (FAMuSS) [3], funded by the National Institutes of Health (NIH), measured a variety of demographic, phenotypic, and genetic characteristics of about 1300 participants. Data from the study has been used in many publications [4]. This section describes data similar to that used in a study of the relationship between muscle strength and a location on the gene *actn3* [5] The study also measured the response in strength of the non-dominant arm to 12 weeks of training.

[3] Thompson PD, Moyna M, Seip, R, et al., 2004. Functional Polymorphisms Associated with Human Muscle Size and Strength. Medicine and Science in Sports and Exercise 36:1132 - 1139

[4] Pescatello L, et al. Highlights from the functional single nucleotide polymorphisms associated with human muscle size and strength or FAMuSS study, BioMed Research International 2013.

[5] Clarkson P, et al., Journal of Applied Physiology 99: 154163, 2005.

Table **??** displays rows 1, 2, 3, and 595 of some of the data from the 595 study participants. The complete set of observations will be referred to as the `FAMuSS` data set. Each row in the table shows the `sex`, `age`, `height`, `weight`, along with a genotype at a particular location (`actn3.r577x`), and response to training (`ndrm.ch`) from a single study participant, or each **case**. [6] The columns represent characteristics, called **variables**, for each of the participants.

For example, the first row represents a Caucasian female, 27 years of age, 65 inches tall, weighing 199 pounds, of genotype CC, who increased strength in her nondominant arm by 40% after training. It is important to understand the definitions of variables, even for items that seem obvious. For example, weight in this dataset has been recorded in pounds, rather than on the metric scale more commonly used outside the United States. Definitions of the variables are given in Table **??**

	sex	age	race	height	weight	actn3.r577x	ndrm.ch
1	Female	27	Caucasian	65.0	199.0	CC	40.0
2	Male	36	Caucasian	71.7	189.0	CT	25.0
3	Female	24	Caucasian	65.0	134.0	CT	40.0
595	Female	30	Caucasian	64.0	134.0	CC	43.8

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

The data in Table 1.4 are organized as a **data matrix**. 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 8, in which the cases were patients and three variables were recorded for each patient.

Data matrices are a convenient way to record and store data. If data is collected for another individual, another row can easily be added; similarly, another column can be added for a new variable.

1.2.2 Types of variables

The `county`[7] dataset summarizes information about the 3,143 counties in the United States. This dataset includes information about each county: its name, the state where it is located, its population in 2000 and 2010, per capita federal spending, poverty rate, and five additional characteristics. Seven rows of the `county` data set are shown in Table 1.5, and the variables are summarized in Table 1.6.

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

[7] Data source: US Bureau of the Census. These data were collected from the US Census website. http://quickfacts.census.gov/qfd/index.html

variable	**description**
`sex`	Sex of the participant
`age`	Age in years
`race`	Recorded as African AM (African American), Am Indian (American Indian), Caucasian, Hispanic and Other
`height`	Height in inches
`weight`	Weight in pounds
`actn3.r577x`	Genotype at the location r577x in the gene actn3. The three genotypes observed were CC, CT and TT
`ndrm.ch`	Percent change in strength in the non-dominant arm, comparing strength after training to strength before training

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

	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.

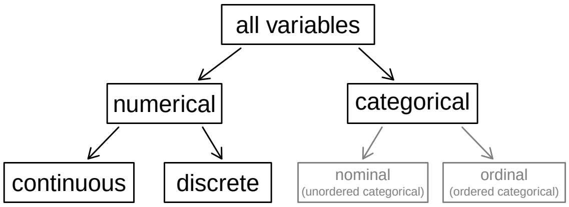


Figure 1.7: Breakdown of variables into their respective types.

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

The variable `fed_spend` is a **numerical** variable since it can take on a wide range of numerical values, and it is possible to add, subtract, or take averages with those values. However, we would not classify a variable reporting telephone area codes as numerical; averages, sums, or differences done with area codes have no clear meaning.

The `pop2010` variable is also numerical, although it is a little different than `fed_spend`. Population count can only be whole, non-negative numbers (`0`, `1`, `2`, ...), so the population variable is said to be **discrete**. 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,[8] and the possible values are called the variable's **levels**.

Finally, the `smoking_ban` variable describes the type of county-wide smoking ban and takes values `none`, `partial`, or `comprehensive` in each county. This variable is a hybrid: it is a categorical variable, but the levels have a natural ordering. A variable with these properties is called an **ordinal categorical** variable. To simplify analyses, ordinal variables in this book will be treated as categorical variables.

● **Example 1.1** Suppose 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 1.2** Consider the variables `group` and `outcome` (at 30 days) from the stent study in Section 1.1. Are these numerical or categorical variables?[9]

1.2.3 Relationships between variables

Many studies are motivated by a researcher examining a possible relationship between two or more variables. In the `county` dataset, a social or public health scientist might be

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

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

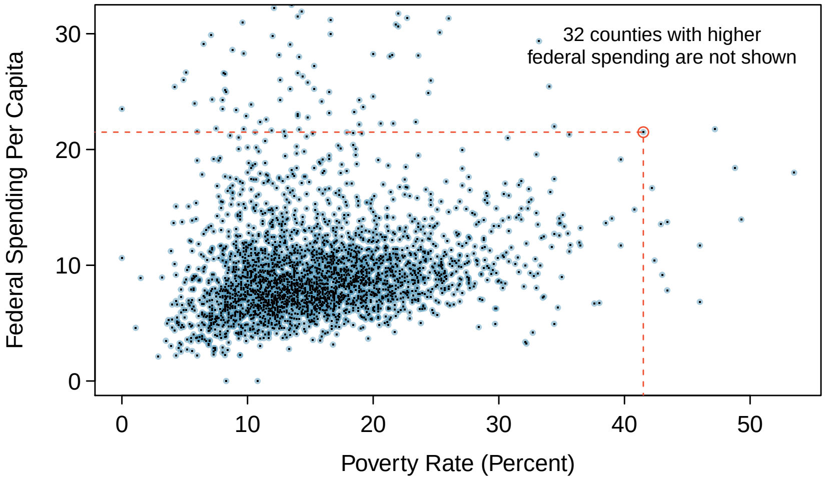


Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

interested 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 percentage 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?

Examining summary statistics may provide some initial insights towards answering these three questions. Graphs are useful for answering such questions as well, as they can be used to visually summarize data.

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 scatterplot suggests a relationship between the two variables: counties with a high poverty rate also tend to have slightly more federal spending.

⊙ **Guided Practice 1.3** Examine the variables in the `FAMuSS` dataset, which are described in Table **??**. Create two questions about the relationships between these variables that could be feasibly addressed in a study.[10]

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.

[10] Two sample questions: (1) Do participants appear to respond differently to training according to genotype? (2) Do male participants appear to respond differently to training than females?

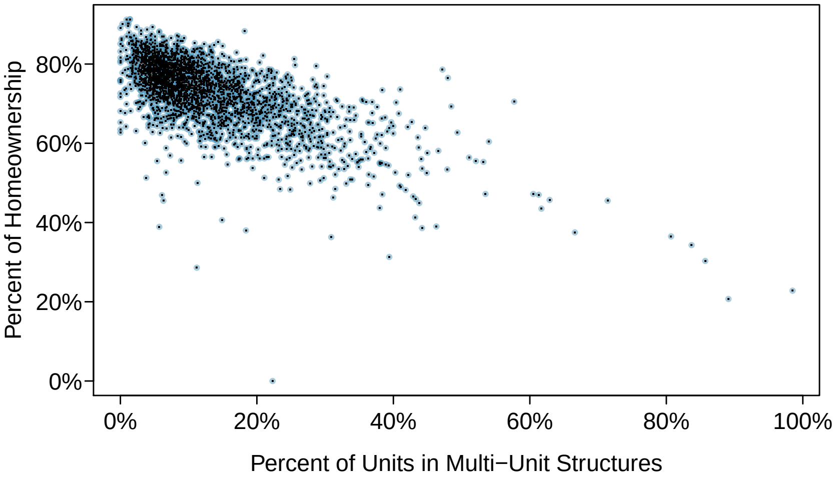


Figure 1.9: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties. Interested readers may find an image of this plot with an additional third variable, county population, presented at www.openintro.org/stat/down/MHP.png.

● **Example 1.4** The relationship between homeownership 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, in which 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.

1.3 Overview of data collection principles

The first step in conducting research is to identify questions to investigate. A clearly articulated research question is essential in identifying which subjects should be studied, what variables are relevant, and how data should be measured. In order to obtain reliable data, it is also important to consider *how* data are collected.

1.3.1 Populations and samples

1. What is the average mercury content of swordfish in the Atlantic Ocean?
2. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each of these questions 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 often 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; this sample data can be used to provide an estimate of the population average in order to answer the research question.

Guided Practice 1.5 For the second question above, identify the target population and what represents an individual case.[11]

1.3.2 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. 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.

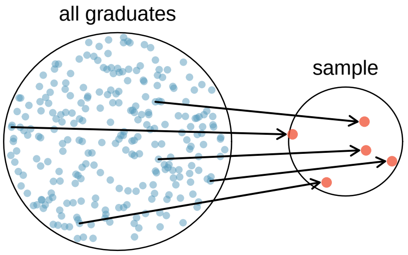


Figure 1.10: In this graphic, five graduates are randomly selected from the population to be included in the sample.

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

Example 1.6 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.

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**

[11] (2) A single person with severe heart disease represents a case. The population includes all people with severe heart disease.

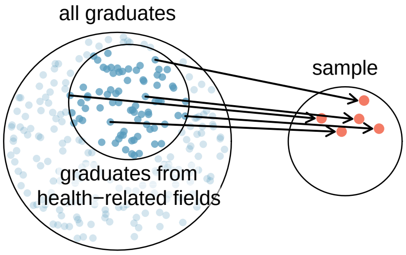


Figure 1.11: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

random sample, and 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, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.

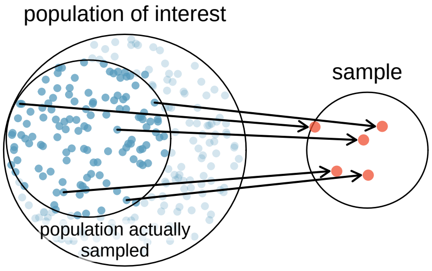


Figure 1.12: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

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

⊙ **Guided Practice 1.7** 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?[12]

[12] 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.

1.3.3 Explanatory and response variables

Consider the following question from page 14 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.[13] 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 and plan an appropriate analysis.

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

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 14:

(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.

1.3.4 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 the data that arise. 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

[13] 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.

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

1.4 Observational studies and sampling strategies

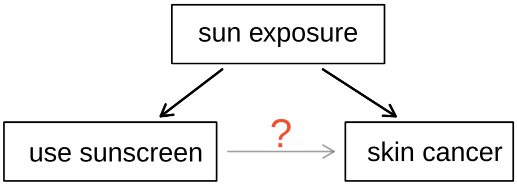
1.4.1 Observational studies

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 1.8** 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?[14]

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.



Sun exposure is what is called a **confounding variable**,[15] 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.

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

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

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 1.9** 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 visible in Figure 1.9.[16]

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.[17] This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets, such as `county`, may contain both prospectively- and retrospectively-collected variables. Local governments prospectively collect some variables as events unfolded (e.g. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

1.4.2 Four 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, these statistical methods – the estimates and errors associated with the estimates – are not reliable. Here we consider four random sampling techniques: simple, stratified, cluster, and multistage sampling. Figures 1.13 and 1.14 provide graphical representations 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 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, since some teams have a lot more money (up to 4 times as much!). Then we might randomly sample 4 players from each team for a total of 120 players.

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

[17] www.channing.harvard.edu/nhs

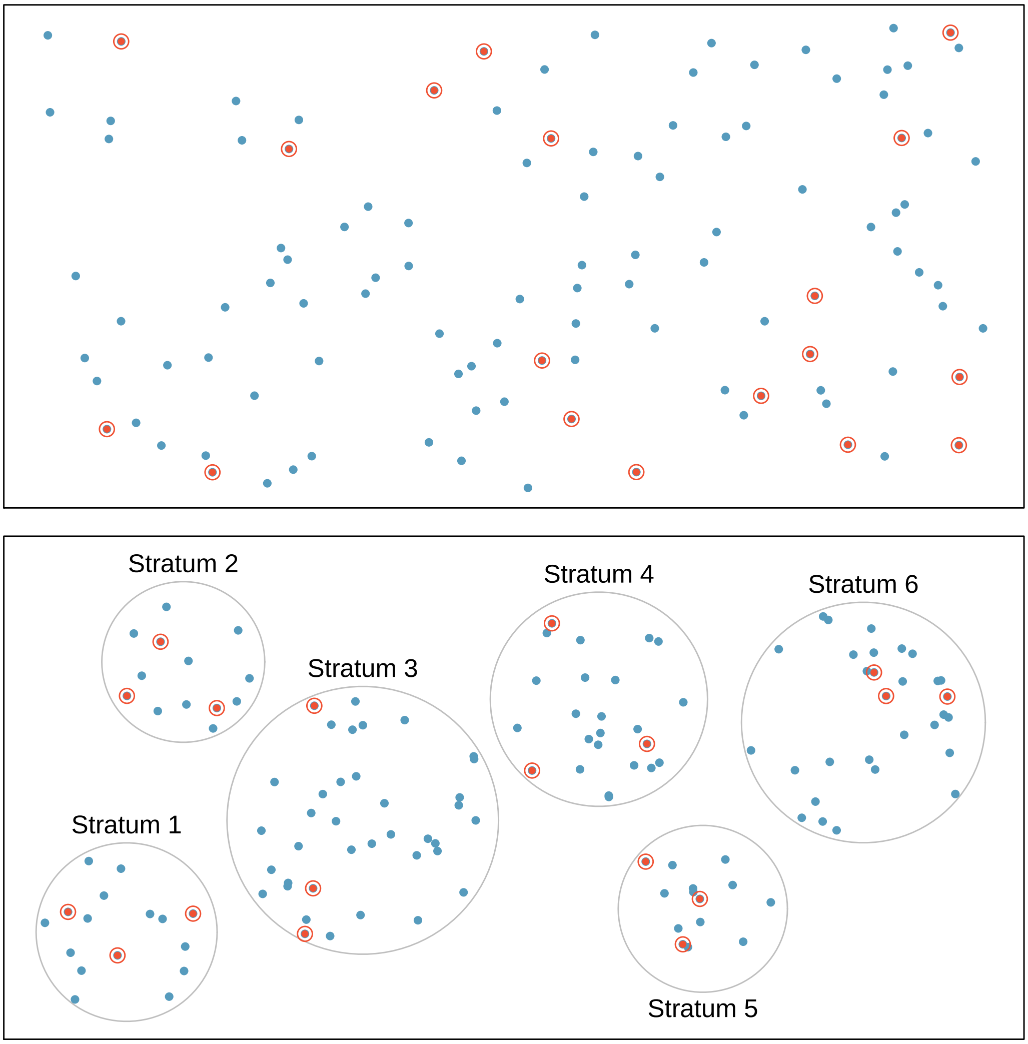


Figure 1.13: Examples of simple random and stratified sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the bottom panel, stratified sampling was used: cases were grouped into strata, then simple random sampling was employed within each stratum.

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 1.10** 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 a **cluster sample**, we break up the population into many groups, called **clusters**. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample. A **multistage sample** is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.

Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques. Also, unlike stratified sampling, these approaches are 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 cluster or multistage sampling work best when the neighborhoods are very diverse. A downside of these methods is that more advanced analysis techniques are typically required, though the methods in this book can be extended to handle such data.

● **Example 1.11** Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

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

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.

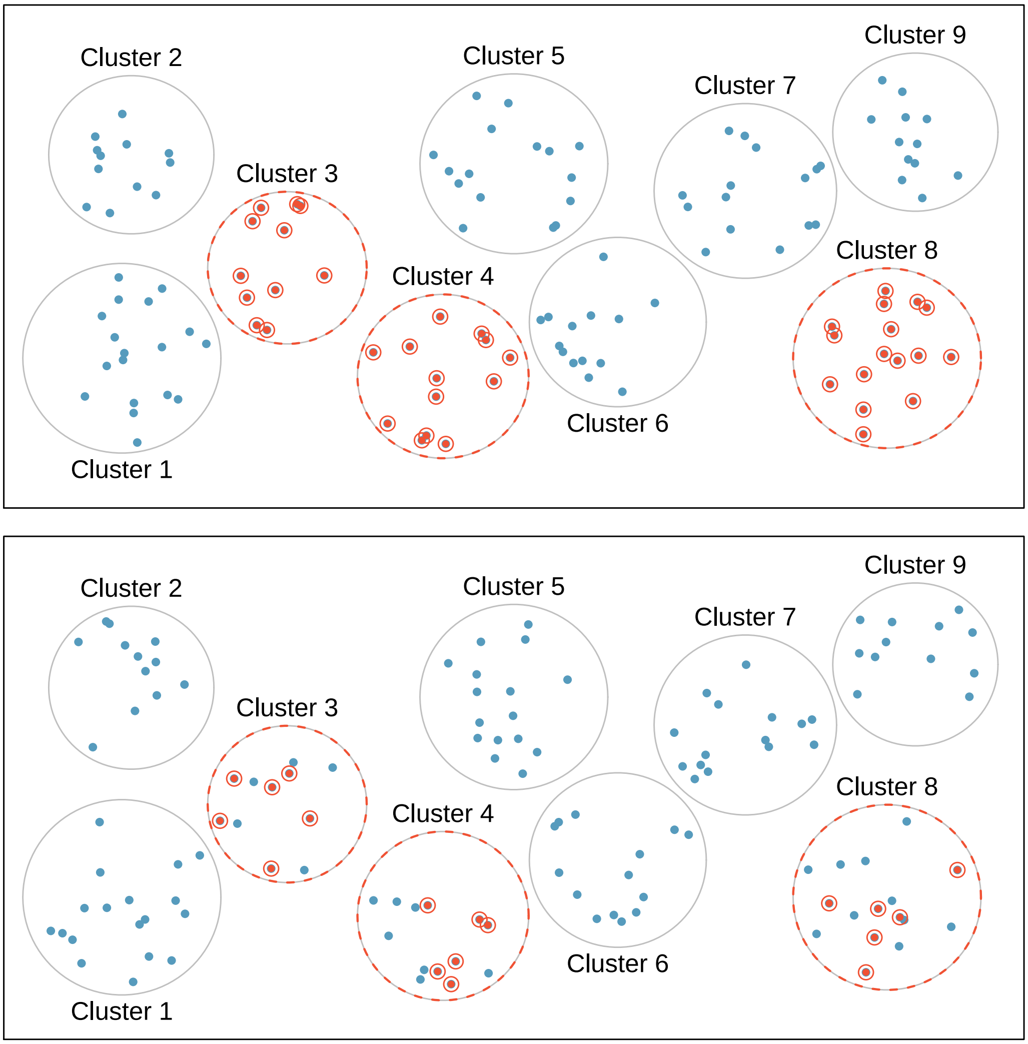


Figure 1.14: Examples of cluster and multistage sampling. In the top panel, cluster sampling was used. Here, data were binned into nine clusters, three of these clusters were sampled, and all observations within these three cluster were included in the sample. In the bottom panel, multistage sampling was used. It differs from cluster sampling in that of the clusters selected, we randomly select a subset of each cluster to be included in the sample.

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 into **blocks** 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 in the study 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 applicable 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.[18] In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[19] 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:

[18] 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.

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

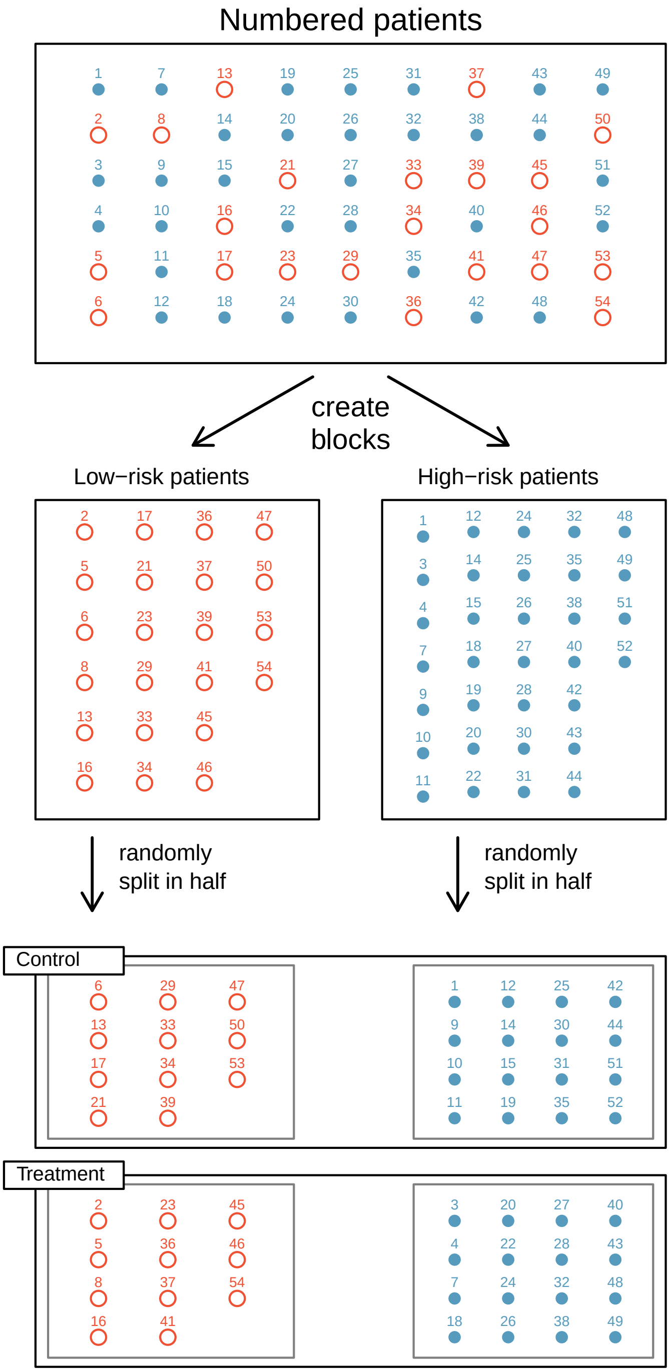


Figure 1.15: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both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

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.[20]

⊙ **Guided Practice 1.12** 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?[21]

1.6 Examining numerical data

In this section we will be introduced to techniques for exploring and summarizing numerical variables. 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, but rather they are categorical variables.

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 14, 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.

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 verbose! Upon further

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

[21] 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.

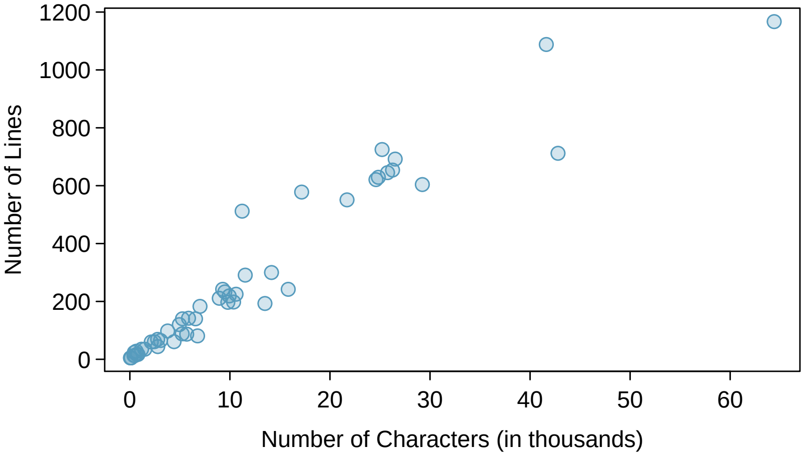


Figure 1.16: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

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 1.13** What do scatterplots reveal about the data, and how might they be useful?[22]

● **Example 1.14** Consider a new data set of 54 cars with two variables: vehicle price and weight.[23] 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 14 and Figure 1.16, which show relationships that are very linear.

⊙ **Guided Practice 1.15** Describe two variables that would have a horseshoe shaped association in a scatterplot.[24]

1.6.2 Dot plots and the mean

Sometimes two variables are 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.

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

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

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

[24] 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.

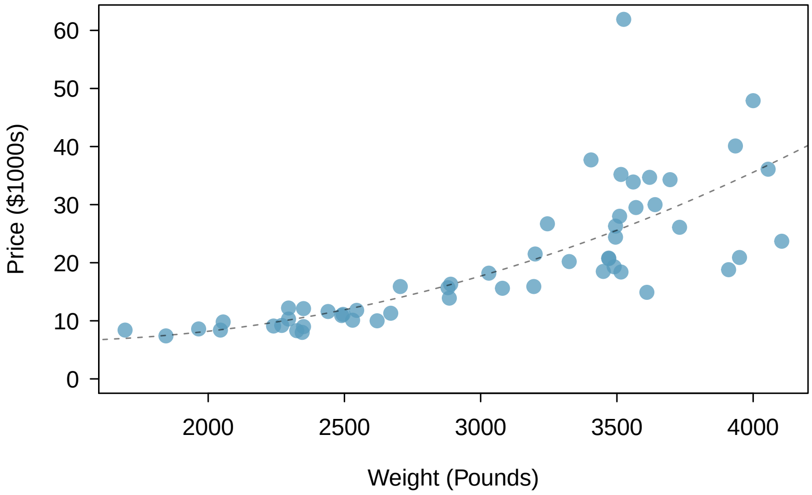


Figure 1.17: A scatterplot of `price` versus `weight` for 54 cars.

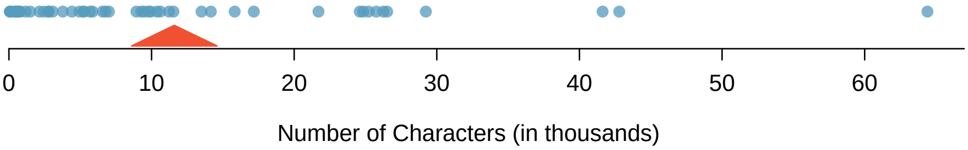


Figure 1.18: A dot plot of `num_char` for the `email50` data set.

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.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \tag{1.16}$$

$\bar{x}$
sample mean

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `num_char`, and the bar over on the x communicates that the average number of characters in the 50 emails was 11,600. It is useful to think of the mean as the balancing point of the distribution. The sample mean is shown as a triangle in Figures 1.18 and 1.19.

Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.17}$$

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

n
sample size

⊙ **Guided Practice 1.18** Examine Equations (1.16) and (1.17) above. What does

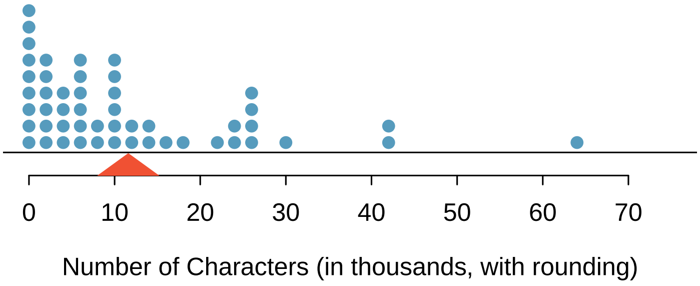


Figure 1.19: A stacked dot plot of `num_char` for the `email50` data set. The values have been rounded to the nearest 2,000 in this plot.

x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[25]

⊙ **Guided Practice 1.19** What was n in this sample of emails?[26]

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

μ
population
mean

● **Example 1.20** 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 4 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.

● **Example 1.21** We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,143 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

[25] 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.

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

$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 1.21 used what is called a **weighted mean**, which will not be a key topic in this textbook. However, we have provided an online supplement on weighted means for interested readers:

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

1.6.3 Histograms and shape

Dot plots show the exact value for 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, we prefer to think of the value as belonging to a *bin*. For example, in the `email50` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. 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.

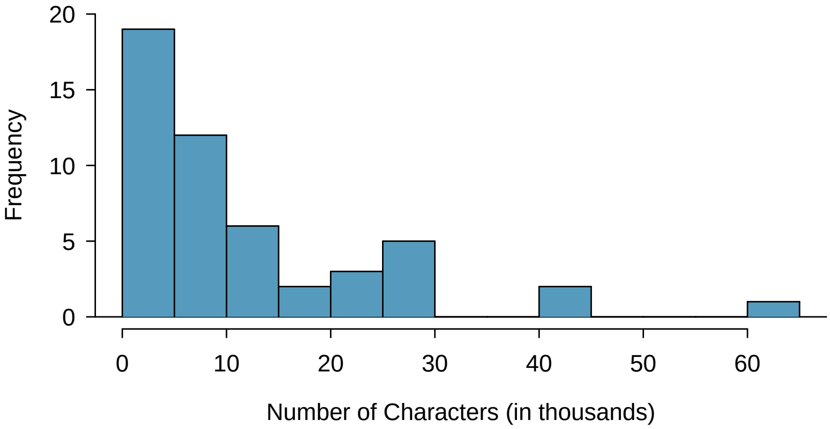


Figure 1.21: A histogram of `num_char`. This distribution is very strongly skewed to the right.

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

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.**[27]

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

Long tails to identify skew

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

⊙ **Guided Practice 1.22** 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?[28]

⊙ **Guided Practice 1.23** Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[29]

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.[30] 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 1.24** Figure 1.21 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[31]

⊙ **Guided Practice 1.25** 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?[32]

[27] 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**.

[28] 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.

[29] Character counts for individual emails.

[30] 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.

[31] 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.)

[32] 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.

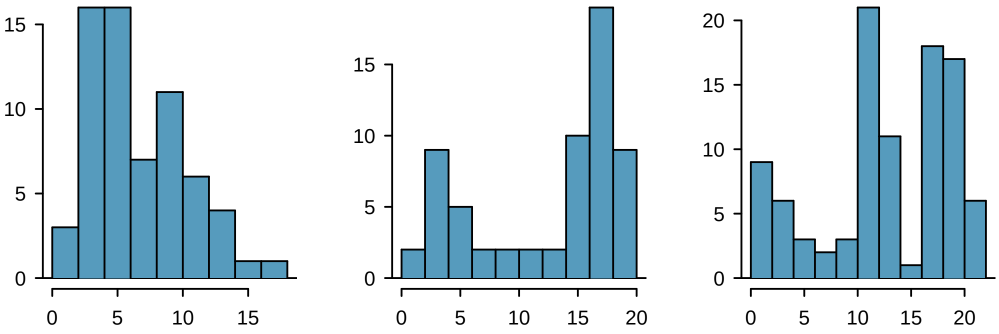


Figure 1.22: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

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 was introduced as a method 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 their formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to 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.

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

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:

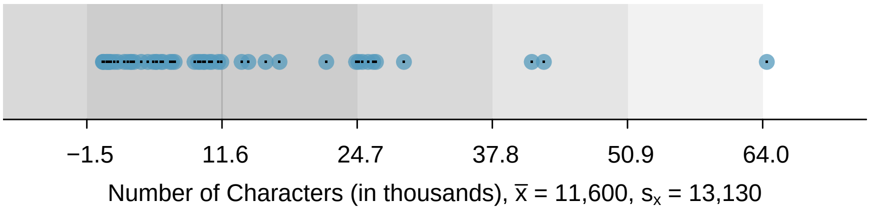


Figure 1.23: In the num_char data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule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

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

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 defined as the square root of the variance:

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

s
sample
standard
deviation

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. The standard deviation is useful when considering 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.[33] 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

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

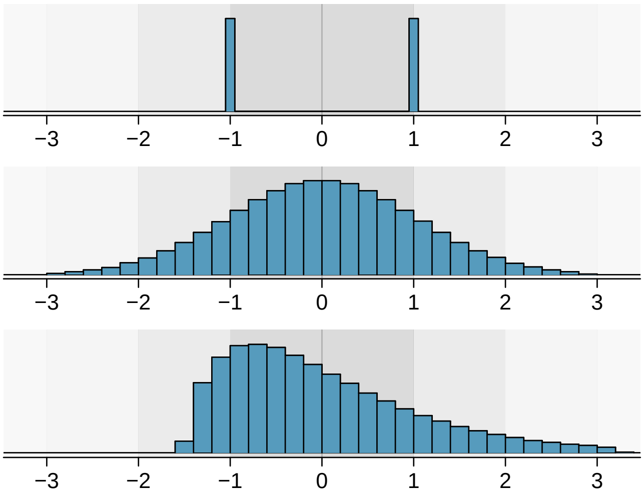


Figure 1.24: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

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 1.26** On page 30, 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.[34]

● **Example 1.27** Describe the distribution of the `num_char` variable using the histogram in Figure 1.21 on page 30. 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 4 we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

[34] 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.

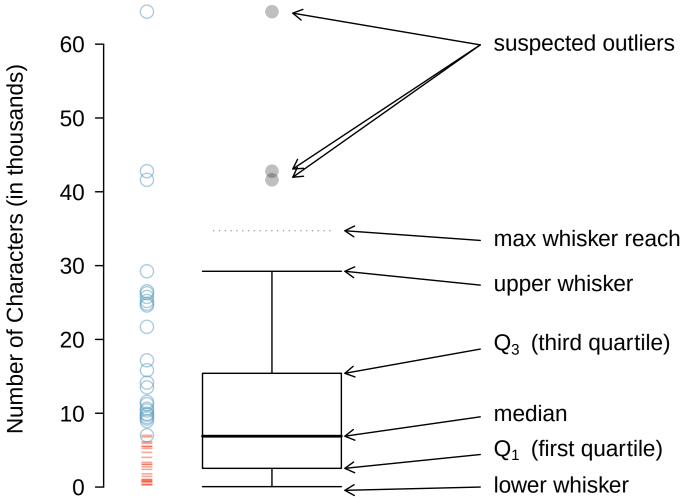


Figure 1.25: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

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 1.28** What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[35]

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 $1.5 \times IQR$.[36] They capture everything within this reach. In Figure 1.25, the upper whisker does not extend to the last three points, which is 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 appears 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 1.29** The observation 64,401, a suspected outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[37]

[35] 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.

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

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

Guided Practice 1.30 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.[38]

Calculator videos

Videos covering how to create statistical summaries and box plots using TI and Casio graphing calculators are available at openintro.org/videos.

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.

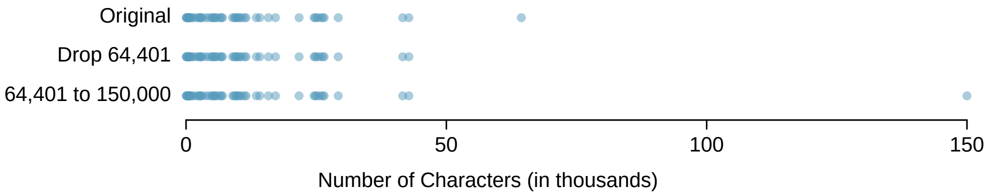


Figure 1.26: Dot plots of the original character count data and two modified data sets.

	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 1.31 (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?[39]

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.

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

[39] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 1.31.

● **Example 1.32** 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 1.33** 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?[40]

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).

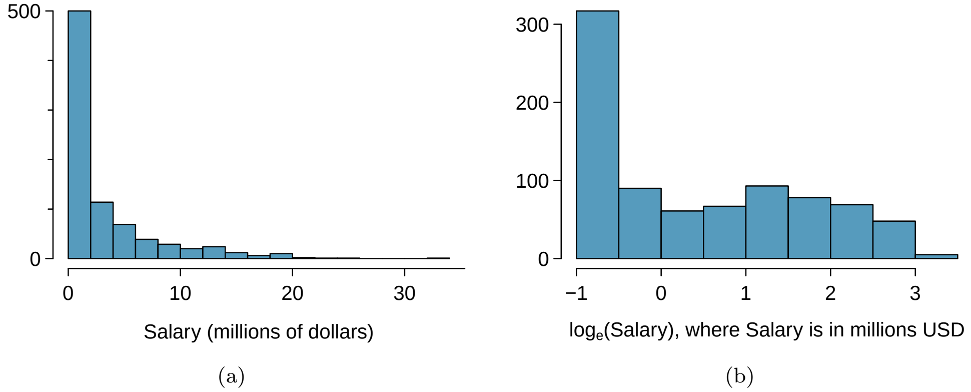


Figure 1.28: (a) Histogram of MLB player salaries for 2010, in millions of dollars. (b) Histogram of the log-transformed MLB player salaries for 2010.

● **Example 1.34** 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[41] of player salaries results in a new histogram in Figure 1.28(b).

[40] 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.

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

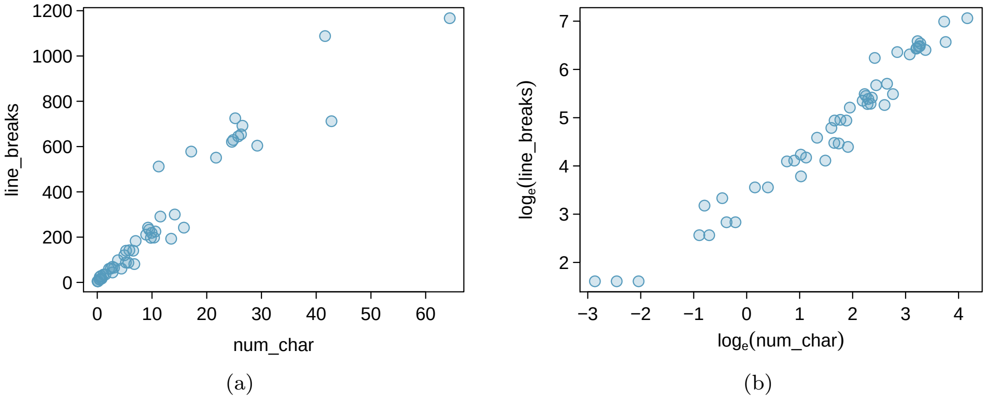


Figure 1.29: (a) Scatterplot of `line_breaks` against `num_char` for 50 emails. (b) A scatterplot of the same data but where each variable has been log-transformed.

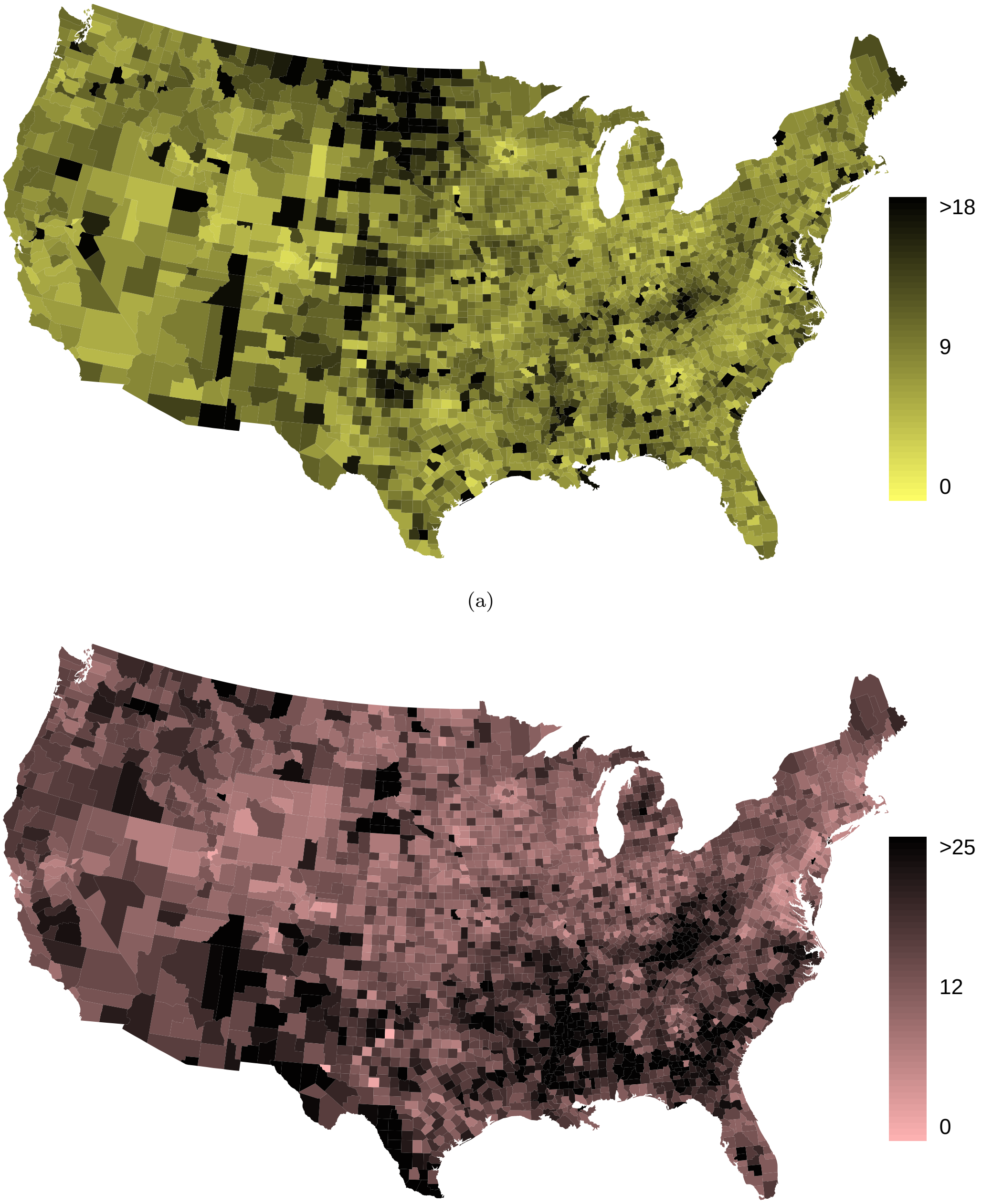
Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

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 **??**, 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.

1.6.8 Mapping data (special topic)

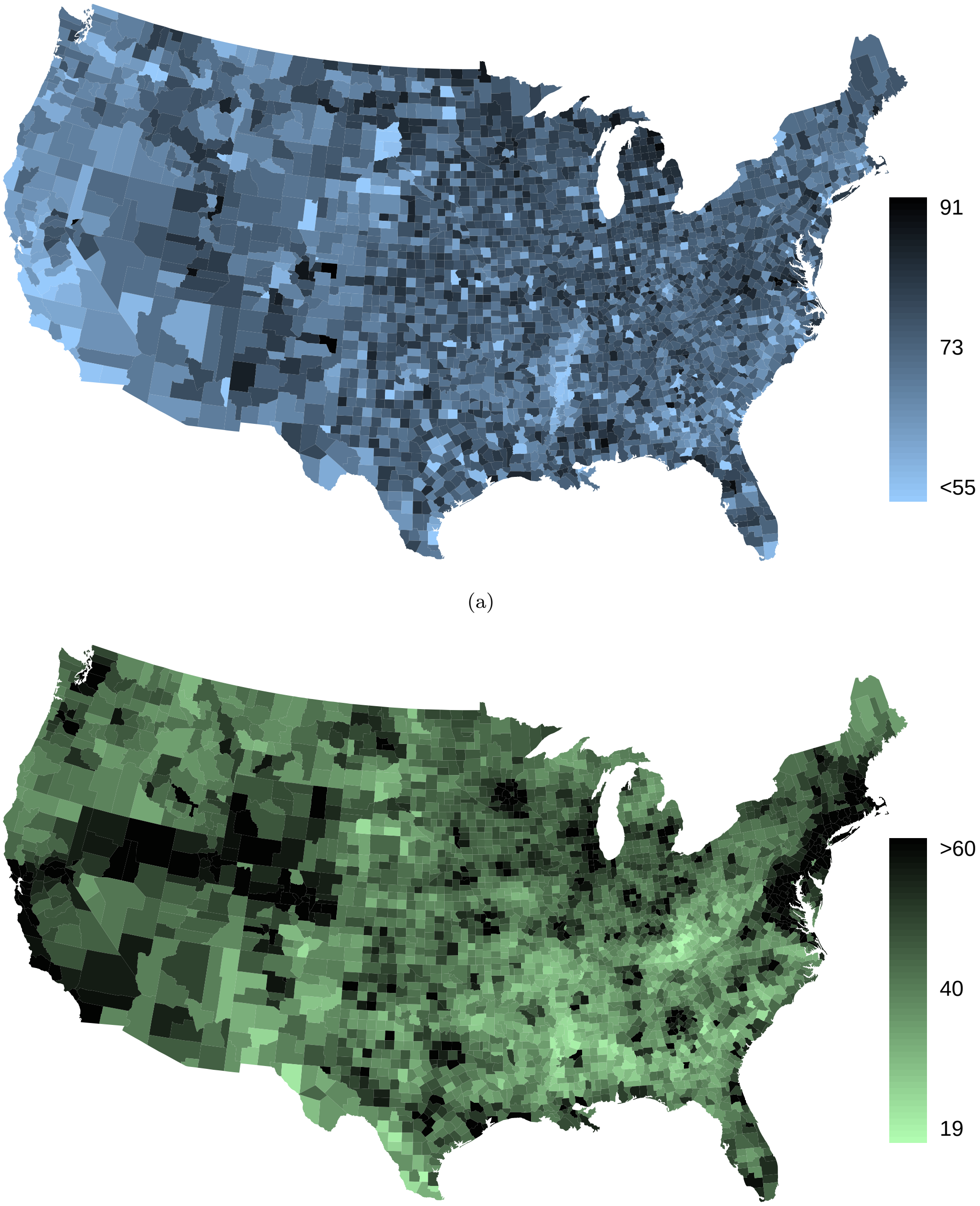
The `county` data set offers many numerical variables that we could plot using dot plots, scatterplots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should map it using an **intensity map**, where colors are used to show higher and lower values 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.



(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).

● **Example 1.35** 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.

⊙ **Guided Practice 1.36** What interesting features are evident in the `med_income` intensity map in Figure 1.31(b)?[42]

[42] 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.

1.7 Considering categorical data

Like numerical data, categorical data can also be organized and analyzed. In this section, we will introduce 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`), showing the proportion of observations that are in each level (i.e. in each category).

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.

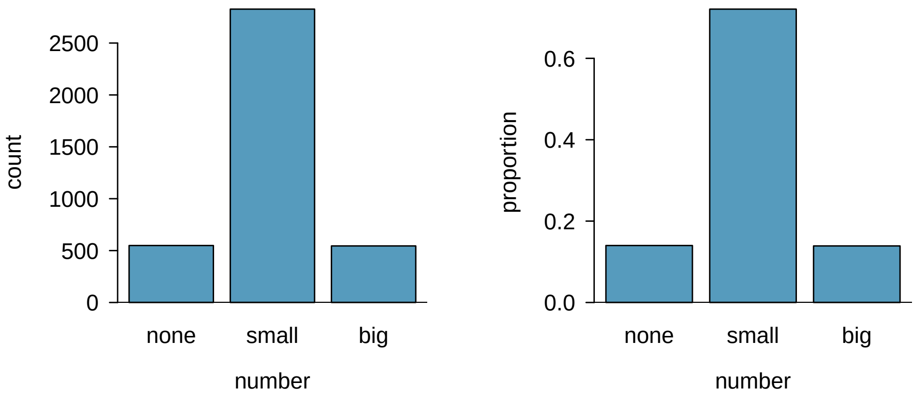


Figure 1.34: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

	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 compared to 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 1.37** What does 0.458 represent in Table 1.35? What does 0.059 represent in Table 1.36?[43]

[43] 0.458 represents the proportion of spam emails that had a small number. 0.059 represents the fraction of emails with small numbers that are spam.

○ **Guided Practice 1.38** 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?[44]

● **Example 1.39** 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 **??**.

	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 1.39 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 1.40** 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?[45]

[44] 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.

[45] 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.

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 divided 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`.

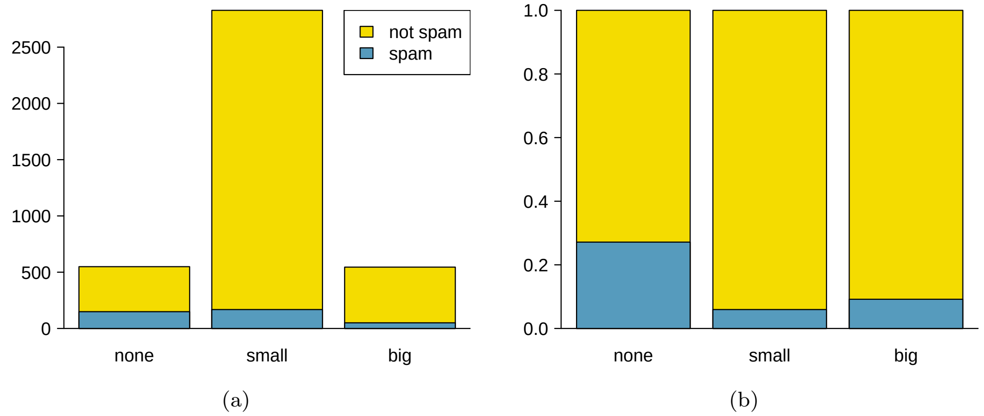


Figure 1.38: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by `spam`. (b) Standardized version of Figure (a).

Example 1.41 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 for each number type.

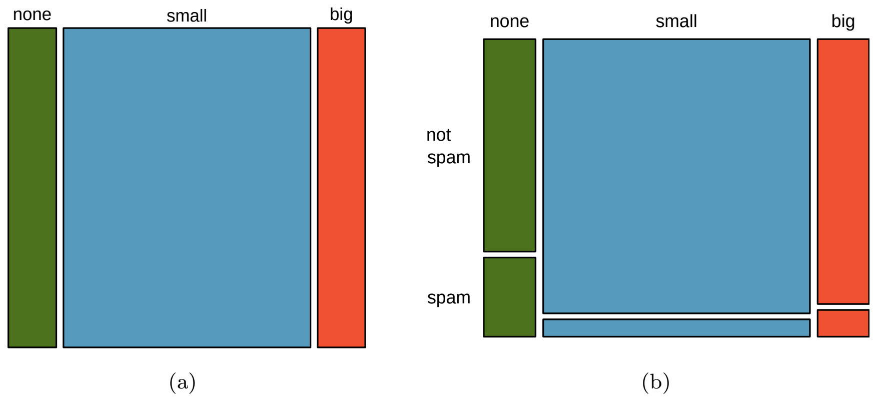


Figure 1.39: The one-variable mosaic plot for `number` and the two-variable mosaic plot for both `number` and `spam`.

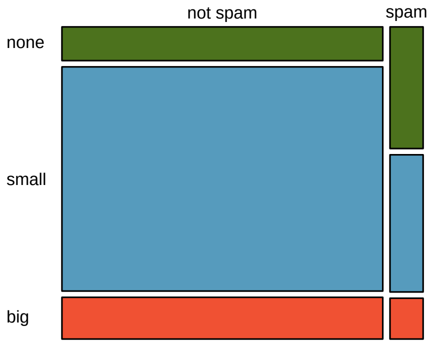


Figure 1.40: Mosaic plot where emails are grouped by the `number` variable after they've been divided into `spam` and `not spam`.

For instance, there are fewer emails with no numbers than emails with only small numbers, so the no number email column is slimmer. In general, mosaic plots use box *areas* to represent the number of observations that box represents.

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).

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!

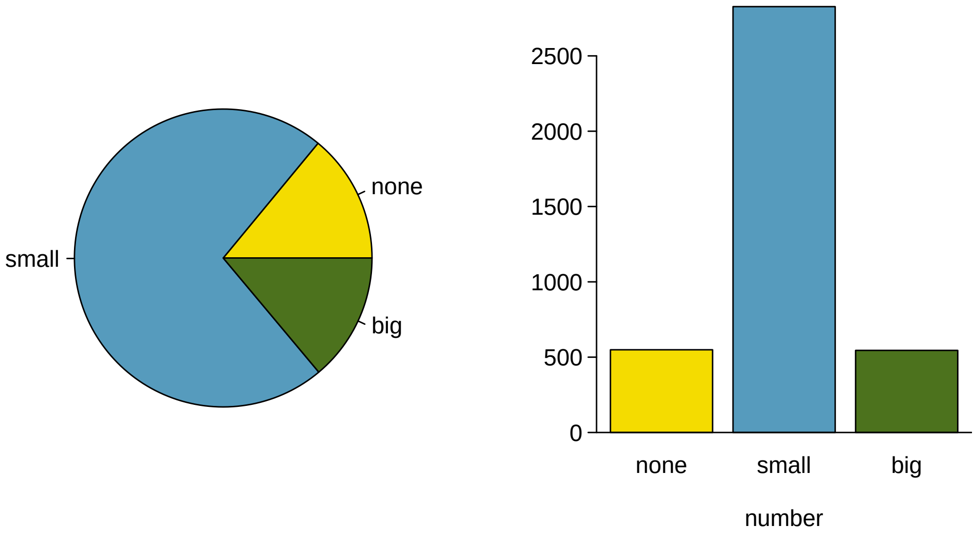


Figure 1.41: A pie chart and bar plot of `number` for the `email` data set.

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					
41.2	33.1	30.4	37.3	79.1	34.5
22.9	39.9	31.4	45.1	50.6	59.4
47.9	36.4	42.2	43.2	31.8	36.9
50.1	27.3	37.5	53.5	26.1	57.2
57.4	42.6	40.6	48.8	28.1	29.4
43.8	26	33.8	35.7	38.5	42.3
41.3	40.5	68.3	31	46.7	30.5
68.3	48.3	38.7	62	37.6	32.2
42.6	53.6	50.7	35.1	30.6	56.8
66.4	41.4	34.3	38.9	37.3	41.7
51.9	83.3	46.3	48.4	40.8	42.6
44.5	34	48.7	45.2	34.7	32.2
39.4	38.6	40	57.3	45.2	33.1
43.8	71.7	45.1	32.2	63.3	54.7
71.3	36.3	36.4	41	37	66.7
50.2	45.8	45.7	60.2	53.1	
35.8	40.4	51.5	66.4	36.1	

no gain		
40.3	33.5	34.8
29.5	31.8	41.3
28	39.1	42.8
38.1	39.5	22.3
43.3	37.5	47.1
43.7	36.7	36
35.8	38.7	39.8
46	42.3	48.2
38.6	31.9	31.1
37.6	29.3	30.1
57.5	32.6	31.1
46.2	26.5	40.1
38.4	46.7	25.9
36.4	41.5	45.7
39.7	37	37.7
21.4	29.3	50.1
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.

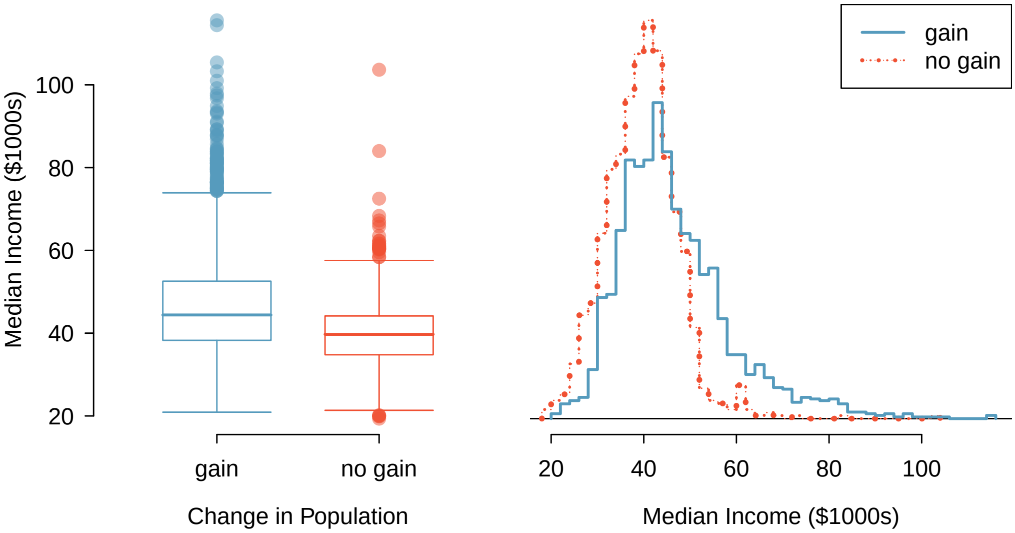


Figure 1.43: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.

⊙ **Guided Practice 1.42** 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?[46]

⊙ **Guided Practice 1.43** What components of each plot in Figure 1.43 do you find most useful?[47]

1.8 Case study: gender discrimination (special topic)

● **Example 1.44** Suppose your professor splits the students in class into two groups: students on the left and students on the right. If $\hat{p}_L$ and $\hat{p}_R$ represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if $\hat{p}_L$ did not exactly equal $\hat{p}_R$?

While the proportions would probably be close to each other, it would be unusual for them to be exactly the same. We would probably observe a small difference due to chance.

⊙ **Guided Practice 1.45** 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?[48]

1.8.1 Variability within data

We consider a study investigating gender discrimination in the 1970s, which is set in the context of personnel decisions within a bank.[49] The research question we hope to answer is, "Are females unfairly discriminated against in promotion decisions made by male managers?"

The participants in this study are 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.

[46] 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.

[47] 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.

[48] We would be assuming that these two variables are independent.

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

Guided Practice 1.46 Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[50]

For each supervisor we record the gender associated with the assigned file and the promotion decision. Using the results of the study summarized in Table 1.44, 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 1.44: Summary results for the gender discrimination study.

Example 1.47 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, what comes to mind as we try to determine whether the data show convincing evidence of a real difference?

The observed promotion rates (58.3% for females versus 87.5% for males) suggest there might be discrimination against women in promotion decisions. However, we cannot 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 1.47 is a reminder that the observed outcomes in the sample may not perfectly reflect the true relationships between variables in the underlying population. Table 1.44 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 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: **Independence model.** 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 model.** 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.

What would it mean if the independence model, which says the variables gender and decision are unrelated, is true? It would mean each banker was going to decide whether to promote the candidate without regard to the gender indicated on the file. That is,

[50] 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.

the difference in the promotion percentages was 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 model: 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 decisions for female personnel files relative to the male files.

We choose between these two competing claims by assessing if the data conflict so much with H_0 that the independence model cannot be deemed reasonable. If this is the case, and the data support H_A, then we will reject the notion of independence and conclude there was discrimination.

1.8.2 Simulating the study

Table 1.44 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 arrangement 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.

In this **simulation**, we thoroughly shuffle 48 personnel files, 24 labeled `male_sim` and 24 labeled `female_sim`, 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_sim` and `female_sim` who were promoted. The randomization of files in this simulation is independent of the promotion decisions, which means any difference in the two fractions is entirely due to chance. Table 1.45 show the results of such a simulation.

		decision		
		promoted	not promoted	Total
gender_sim	male_sim	18	6	24
	female_sim	17	7	24
	Total	35	13	48

Table 1.45: Simulation results, where any difference in promotion rates between `male_sim` and `female_sim` is purely due to chance.

⊙ **Guided Practice 1.48** What is the difference in promotion rates between the two simulated groups in Table 1.45? How does this compare to the observed 29.2% in the actual groups?[51]

[51] $18/24 - 17/24 = 0.042$ 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.

1.8.3 Checking for independence

We computed one possible difference under the independence model in Guided Practice 1.48, which represents one difference due to chance. While in this first simulation, we physically dealt out files, it is 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 1.46 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 that were recommended for promotion.

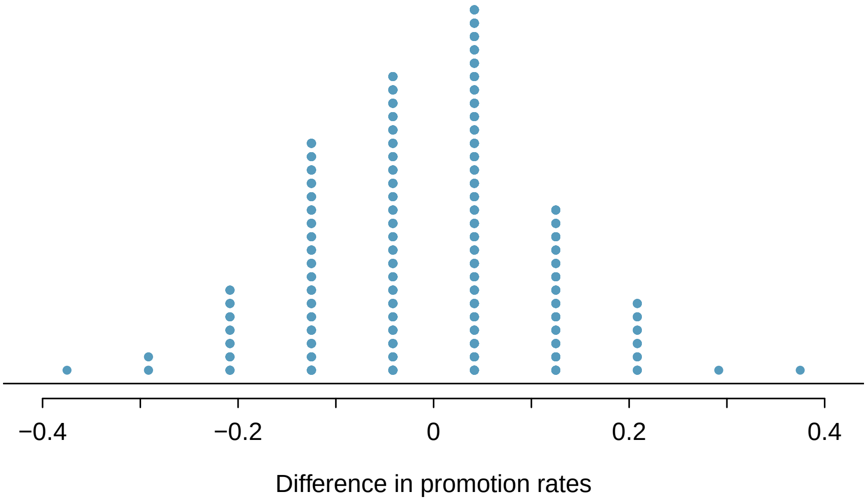


Figure 1.46: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where `gender_sim` and `decision` are independent. Two of the 100 simulations had a difference of at least 29.2%, the difference observed in the study.

Note that the distribution of these simulated differences is centered around 0. We simulated these differences assuming that the independence model was true, and under this condition, we expect the difference to be zero with some random fluctuation. We would generally be surprised to see a difference of *exactly* 0: sometimes, just by chance, the difference is higher than 0, and other times it is lower than zero.

● **Example 1.49** How often would you observe a difference of at least 29.2% (0.292) according to Figure 1.46? 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 1.46. Such a low probability indicates a rare event.

The difference of 29.2% being a rare event suggests two possible interpretations of the results of the study:

H_0 **Independence model.** Gender has no effect on promotion decision, and we observed a difference that would only happen rarely.

H_A **Alternative model.** 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%.

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong evidence against the independence model. That is, we do not have sufficiently strong evidence to conclude there was gender discrimination. (2) We conclude the evidence is sufficiently strong to reject H_0 and assert that there was gender discrimination. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event.[52] So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence of gender discrimination against women by the supervisors.

One field of statistics, statistical inference, is built on evaluating whether such differences are due to chance. In statistical inference, statisticians evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur. In Chapter 4, we give a formal introduction to the problem of model selection. We spend the next two chapters building a foundation of probability and theory necessary to make that discussion rigorous.

[52] 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.

Chapter 4

Foundations for inference

Statistical inference takes what we learned in previous chapters about distributions and probabilities and uses these tools to estimate properties or parameters about a larger population using observed datasets. Statistical inference is also concerned with understanding the quality of these parameter estimates. Once we arrive at a parameter estimate, we can ask ourselves how sure or how confident we are that this estimate is representative of the greater population. For example, a classic inferential question is, "How sure are we that the estimated mean, $\bar{x}$, is near the true population mean, μ?" Statistical inference includes testing these questions but also determining which estimates to use.

Chapter 4 provides the groundwork for inference on a larger population and in later chapters, we will even learn how to use a representative dataset to compare two distinct populations. While the equations and details change depending on the setting, the foundations and general procedures for inference are the same throughout statistics. Chapter 4 covers inference on point estimates from one sample while Chapter **??** builds and expands to inferring from two samples. Chapter **??** steps beyond numerical data to inference on categorical data. Understanding the foundation with point estimates in this chapter will provide familiarity for upcoming chapters.

Let's consider the `BRFSS` data from 2000. The Behavioral Risk Factor Surveillance System (BRFSS) by the CDC was started in 1984 and is the world's largest on-going telephone health survey system. This survey is nationwide and aims to "monitor state-level prevalence of major behavioral risks among adults associated with premature morbidity and mortality." Topics like smoking, alcohol use, exercise, diet and other illnesses are included in this questionnaire. The dataset includes records for 184,450 respondents and 289 variables. Take a look at the `BRFSS` dataset and how the dataset is organized in Table ??. We are particularly interested in the 5 variables listed in Table 4.1

variable	**description**
`sex`	Male or Female where 1 is Male
`age`	Age, in years
`height`	In feet and inches where, for example, 5' 5" is listed as 505
`weight`	In pounds
`wtdesire`	How much would you like to weigh [1]

Table 4.1: Variables of interest and their descriptions for the `BRFSS` data set.

age	weight	wtdesire	height	htf	hit	...	sex
61	205	170	506	5	6	...	2
25	201	201	602	6	2	...	1
40	114	114	507	5	7	...	2
36	190	180	600	6	0	...	1
48	150	125	503	5	3	...	2
⋮	⋮	⋮	⋮	⋮	⋮		⋮

Table 4.2: Five observations from the BRFSS data set.

The CDC, through this survey, is hoping to infer on the characteristics of adults in the United States. Therefore the population that we are interested in with this data is all US adults. We assume that our set of 170,000 observations is representative of this population. If we take a sample of 40,000 adults from the population and pretend that we have no observed the other 130,000 observations, we can use use these 40,000, our sample, to make inferences on our target population. A target population is the group that the statistician is interested in and wants to draw conclusions about. In this case, our target population is US adults. Remember, statistical inference is being able to draw conclusions about a larger population, our target population, through estimates and hypothesis testing. What sorts of estimates and conclusions do you think we can make given this dataset given our sample of 40,000?

Of the variables shown above, we are particularly interested in `weight` and `height` which asks "About how much do you weigh without shoes?" and "About how tall are you without shoes?" While a fairly simple question, this is especially useful to address behavioral and lifestyle choices of Americans including diet and exercise. A person's BMI or Body Mass Index has been a helpful tool to capture both a person's height and weight within one measurement. The BMI is also a measure of body fat based on such height and weight and is used to assess what is considered healthy or desirable. In medicine, BMI can be used to categorize a person as "underweight," "overweight" or "obese." The calculation of a BMI index using both the Metric and the Imperial System is

$$BMI = \frac{\text{weight}_{\text{kg}}}{{\text{height}_{\text{m}}}^2} = \frac{\text{weight}_{\text{lb}}}{{\text{height}_{\text{in}}}^2} \cdot 703$$

Before we can continue even exploring the data with tools from Chapter 1, we need to clean the dataset, a step that many statisticians have to do in practice, and calculate each respondents BMI since the `BRFSS` data does not provide those values. The R code to do the cleaning and calculation can be accessed **in the archives**.

Now that we have cleaned up the dataset, assume that we've taken a simple random sample of `BRFSS` to draw conclusions about the entire target population. This dataset called `brfss.sample` comprises of 40,000 respondents of the original 170,000. Let's explore the data with the tools from Chapter 1 before we continue with estimating our parameters in Figure 4.4 with a snapshot of our new data `brfss.sample` in Table 4.3.

These data from `brfss.sample` is special because in order to do statistical inference, the dataset needs to be representative of the population of interest. In this case because the size of both `BRFSS` and `brfss.sample` is so large, we can assume that the potential of outliers will have significantly less impact on our parameter estimations even though there are many outliers. Now that we have a general idea of what the data looks like, we can start with statistical inference.

age	weight	wtdesire	height	sex	height.total	bmi
60	200	150	508	2	68	30.41
25	145	125	506	2	66	23.40
40	180	170	511	1	71	25.10
53	210	175	511	1	71	29.29
80	170	170	504	2	64	29.18
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.3: Five observations from the new `brfss.sample` data set.

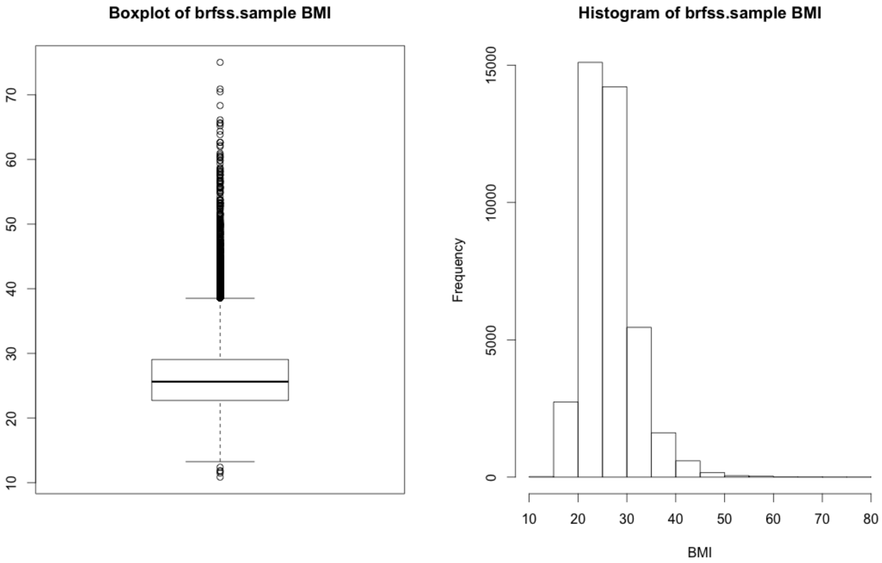


Figure 4.4: Histogram of BMI for the `brfss.sample` data. The data is skewed right by both the histogram and the box plot. We can see from the box plot that it has many outliers

4.1 Variability in estimates

Right now we are interested in the `bmi` variable [2] and want to ask *What is the average BMI of the US population of adults?*.

This question and many others from the `brfss.sample` sample can be useful for monitoring obesity and informative for other behavioral risks. The WHO, World Health Organization, uses this value to categorize individuals while noting that it does not consider muscularity, which affects the BMI value significantly. Below is a categorization of BMI values and ranges.

[2] `bmi` will refer to the variable and BMI will refer to the general measurement

Category	BMI range
Very severely underweight	<15
Severely underweight	15.0-16.0
Underweight	16.0-18.5
Normal (healthy weight)	18.5-25
Overweight	25-30
Obese Class I (Moderately obese)	30-35
Obese Class II (Severely obese)	35-40
Obese Class III (Very severely obese)	40+

For notation, we will use x_1, ..., $x_{40,000}$ to represent the BMI for each survey respondent in our sample `brfss.sample`. [3]

4.1.1 Point estimates

Motivated by our original question, ”What is the average BMI value of the US population of adults?” a likely choice to estimate the **population mean** (the mean BMI of the population) using our sample is to simply take the **sample mean**. That is, to estimate the average BMI of all 40,000 survey respondents in our sample:

$$\bar{x} = \frac{30.40 + 23.40 + 25.10 + \dots}{40,000} = 26.3555$$

The sample mean[4] $\bar{x} = 26.35088$ is called a **point estimate** of the population mean, a single value that is considered to be our best guess if we were only allowed to choose one value to estimate the population mean. Suppose from the original total respondents, we take a new sample of 40,000 people and recompute the mean; we will probably not get the exact same answer that we got using the `brfss.sample` data set. Estimates generally vary from one sample to another, and this **sampling variation** suggests our estimate may be close, but it will not be exactly equal to the parameter, the true average weight of all adults in the United States.

What about generating point estimates of other **population parameters**, such as the population median or population standard deviation? Once again we might estimate parameters based on sample statistics, as shown in Table 4.5. For example, we estimate the population standard deviation for BMI using the sample standard deviation, and the population median using the sample median.

`BMI`	estimate	parameter
mean	26.35551	26.35088
median	25.62044	25.60354
st. dev.	5.288141	5.319992

Table 4.5: Point estimates and parameter values for the `bmi` variable.

⊙ **Guided Practice 4.1** Suppose we want to estimate the average number of pounds that US adults would like to gain or lose. If $\bar{x}_{\text{weight}} = 168.212$ lbs and $\bar{x}_{\text{desired.weight}}$

[3] While we focus on the mean in this chapter, questions regarding variation are often just as important in practice. For instance, potential action regarding obesity could change if the standard deviation of a person’s BMI was 5 versus if it was 15.

[4] If we were interested in the values of another variable, `weight` denoted $w_1, \dots w_{40,000}$ instead, we would denote the sample mean as $\bar{w}$

= 153.199 lbs, then what would be a reasonable point estimate for the population desired change in weight?[5]

⊙ **Guided Practice 4.2** If you had to provide a point estimate of the population IQR for the BMI of participants, how might you make such an estimate using a sample?[6]

4.1.2 Accuracy and Precision of Point Estimates

As we saw above, the sample mean calculated from this sample of 40,000 will likely be different from the sample mean of a different set of 40,000 respondents. Therefore we note that estimates will generally not be exactly the truth, in this instance, the average population BMI. However the accuracy of the point estimate will get better once more data becomes available as we see in Section **??**.

Consider a running mean from the BRFSS data. A **running mean** is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence. In this case, the second mean is the average of the first two observations, x_1, x_2. The third number in the running mean sequence is the average of x_1, x_2, and x_3. The running mean for vmi in the brfss.sample dataset is shown in Figure 4.6. We look at a running mean of 100, 1000, 10000 and 40000 observations. We note that as more values get included, the running mean converges closer to the sample mean of 26.35551 by the plots. The point estimate, the sample mean, becomes a more accurate and precise measurement of the population average as more data becomes available.

Sample point estimates can only approximate the population parameter, and they vary from one sample to another. If we took another simple random sample of 40,000 from all the BRFSS respondents, we would find that the sample mean for BMI would be a little different. It will be useful to quantify how variable an estimate is from one sample to another. If this variability is small (i.e. the sample mean doesn't change much from one sample to another) then that estimate is probably very accurate. If it varies widely from one sample to another, then we should not expect our estimate to be very good. Again, however, we can already get an idea that the point estimate, in this case the sample mean, becomes more precise as well as more data becomes available. **through standard error section**

4.1.3 Standard error of the mean

From the random sample represented in brfss.sample, we estimated the average BMI of an adult in the United States is 26.35551. Suppose we take another random sample of 40,000 individuals and take its mean. We then get 26.35943. Suppose we took another (26.32957) and another (26.39974), and continue to do this many many times – which we can do only because we are sampling from a large, more representative BRFSS dataset[7] – we can build up a **sampling distribution** for the sample mean when the sample size is 40,000, shown in Figure ??.

[5] We could take the difference of the two sample means: 168.212 − 153.199 = 15.01302. US adults on average want to lose 15.01 lbs.

[6] To obtain a point estimate of the IQR for the population, we could take the IQR of the sample.

[7] The sampling distribution depends on the underlying distribution of the population. In this case, while BRFSS is not quite the target population of all US adults, it is large enough to represent the US population as explained above. If we had complete data from the target population, there would be no need to take a sample mean measurement. Normally we aren't even capable of taking another sample of 40,000 from BRFSS!

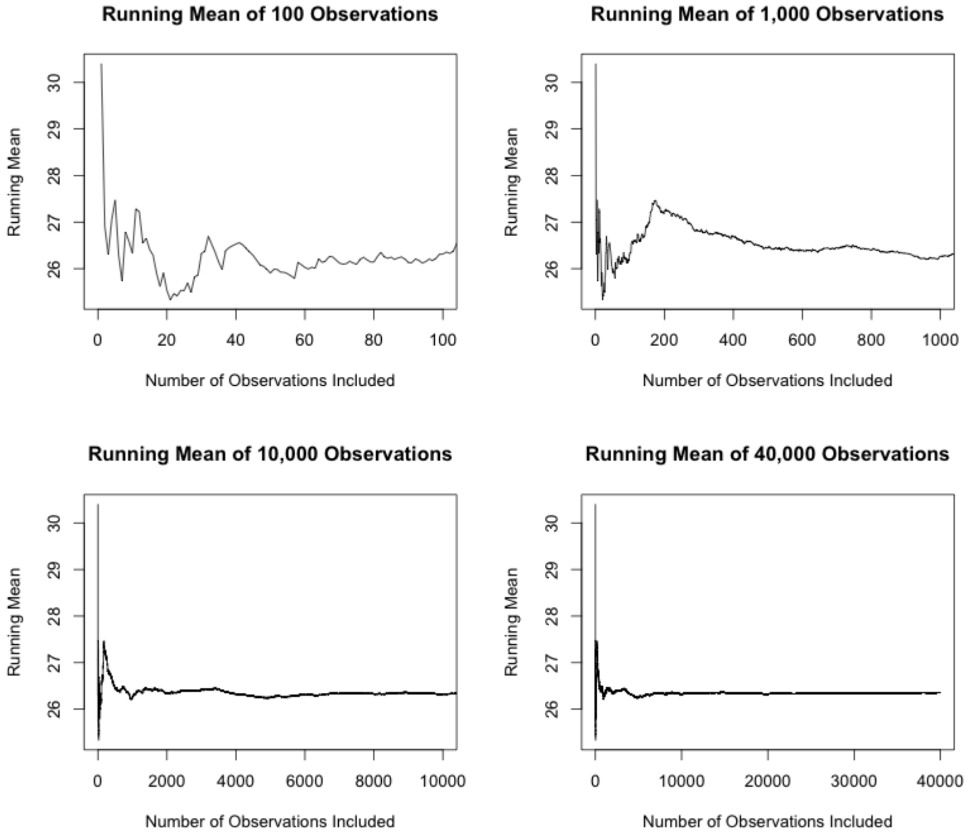


Figure 4.6: The mean computed after adding each individual BMI observation to the sample. The mean tends to approach the true population average as more data become available.

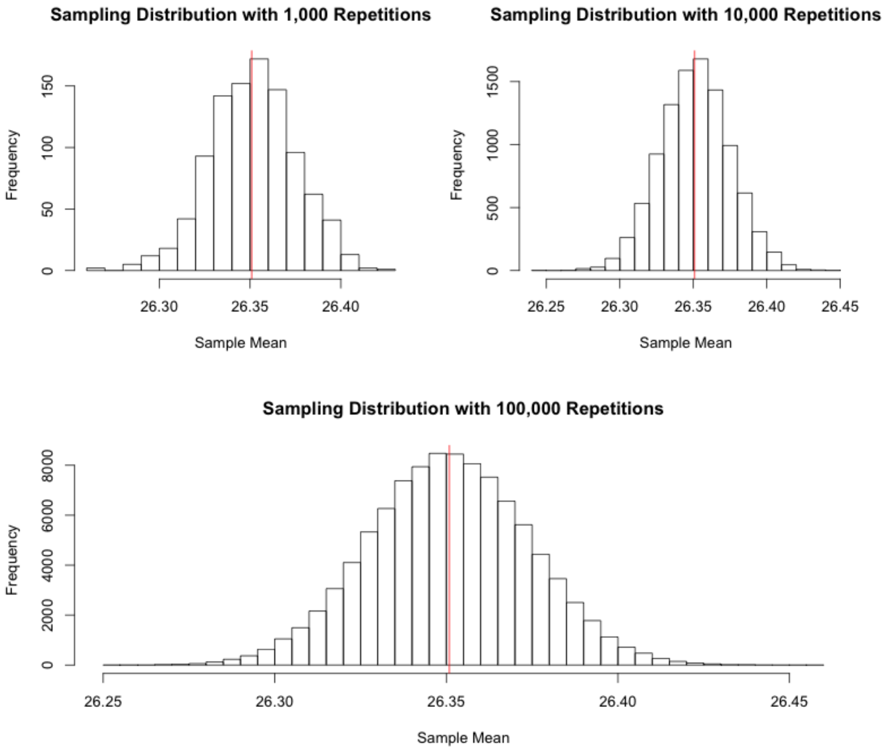


Figure 4.7: A histogram of 1,000, 10,000 and 100,000 sample means for BMI, where the samples are of size $n = 40,000$.

Sampling distribution

The **sampling distribution** of a point estimate represents the distribution of the point estimate based on samples of a fixed size from a certain population. There is a unique sampling distribution that exists that is inherent to the point estimator you are measuring. Every time that you are calculating your point estimate from a particular sample of said size, your point estimate itself is one sample from the sampling distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

Figure ?? is an approximation of the sampling distribution. To truly get the sampling distribution, one would need to sample every possible unique combination of 40,000 respondents from the entire US adult population (and not just the `BRFSS` data set). However we note that just as the running mean becomes a better approximation of the population average as more data becomes available, the approximation of the sampling distribution also becomes more accurate as we take more and more samples as shown by the increasing number of repetitions and the red line at the mean of BMIs in `BRFSS`. We can create an approximation of the sampling distribution of the sampling mean with the following pseudocode [8]:

```
(1) Have a place to store all the sample means that we will calculate
(2) Take a sample from the BRFSS dataset of 40,000
(3) Calculate the sample mean from this specific sample and store it in (1)
(4) Repeat (2) and (3) many many times
(5) Plot all the sample means you have stored in (1) as a histogram
```

As we see above in Figure ?? , as we do more and more repetitions, we get a better approximation of the sampling distribution by the Law of Large Numbers. The sampling distribution, in this case, is likely to be unimodal and approximately symmetric. The sampling distribution is also centered exactly at the true population mean: $\mu = 26.35088$. Intuitively, this makes sense. The sample means should tend to "fall around" the population mean.

By the sampling distribution, we also see that the point estimator will have some variability. Point estimators will vary sample by sample. To quantify this variability of the sample mean around the population mean, we use the standard deviation of the sampling distribution denoted $\sigma_{\bar{x}}$ or s in some contexts. In this book we will continue to use $\sigma_{\bar{x}}$ Just as with the definition of standard deviation in **Chapter 1**, the standard deviation of the sample mean, in this case $\sigma_{\bar{x}} = 0.02659996$, tells us how far the typical estimate is away from the actual population mean. It also is a very good metric for the typical **error** of the point estimate, and for this reason we usually call this version of standard deviation the **standard error (SE)** of the estimate. [9]

SE
standard
error

Standard error of an estimate

The standard deviation associated with an estimate is called the *standard error*. It describes the typical error or uncertainty associated with the estimate.

[8] Refer to the appendix for R code

[9] In general standard error is the standard deviation of samples and estimates whereas we use the term standard deviation for populations or distributions. Look AT SOME REFERENCE for a clearer distinction of standard error versus standard deviation.

The standard error could be calculated if statisticians knew the sampling distribution. It would simply be the standard deviation of that sampling distribution. However when considering the case of the point estimate $\bar{x}$, there is one problem: there is no obvious way to estimate its standard error from a single sample. Statistical theory and computational methods, instead, provide helpful tools to address this issue.

⊙ **Guided Practice 4.3** (a) Would you rather use a small sample or a large sample when estimating a parameter? Why? (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard error than a point estimate based on a larger sample?[10]

In the sample of 40,000 US adults, the standard error of the sample mean is equal to one-two hundredth of the population standard deviation: 5.319992/200 = 0.02659996[11]. In other words, the standard error of the sample mean based on 40,000 observations is equal to

$$SE_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{5.319992}{\sqrt{40000}} = 0.02659996$$

where σ_x is the standard deviation of the individual observations, the population BMIs and n is the number of observations in the sample. This is no coincidence. We can show mathematically that this equation is correct when the observations are independent using the probability tools of Section **??**.

Computing SE for the sample mean

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to

$$SE_{\text{sample mean}} = \frac{\sigma}{\sqrt{n}} \qquad (4.4)$$

A reliable method to ensure sample observations are independent is to guarantee that the sample you have from the population is a simple random sample with a size that is less than 10% of the population.

There is one subtle issue of Equation (4.4) that you might have realized: the population standard deviation is typically unknown. There are two ways to resolve this problem. First, you can use the point estimate of the standard deviation (the standard deviation of the sample) as a replacement for the population standard deviation. This estimate tends to be sufficiently good when the sample size is at least 30 and the population distribution is not strongly skewed. Thus, we often just use the sample standard deviation denoted s instead of σ for the population standard deviation. When the sample size is smaller than 30, we will need to use a method to account for extra uncertainty in the standard error.

[10] (a) Consider two random samples: one of size 10 and one of size 1000. Individual observations in the small sample are highly influential on the estimate while in larger samples these individual observations would more often average each other out. The larger sample would tend to provide a more accurate estimate. (b) If we think an estimate is better, we probably mean it typically has less error. Based on (a), our intuition suggests that a larger sample size corresponds to a smaller standard error.

[11] Again, remember we are assuming that the population standard deviation is the standard deviation of `brfss.df`

If the skew condition is not met, a larger sample is needed to compensate for the extra skew. These topics are further discussed in Section 4.4. For this solution, we would use $s = 5.288141$ and get

$$SE_{estimate} = \frac{5.288141}{\sqrt{40000}} = \frac{5.288141}{200} = 0.02644071$$

A second method to "replace" the population standard deviation would be similar to creating the approximation of the sampling distribution but rather, you can repeatedly sample from `BRFSS` and calculate the sample mean each time. Once you finish with your iterations, calculate the standard deviation of all of these sample means to get an estimation of the Standard Error of your point estimate. In pseudocode:

```
(1) Have a place to store all the sample means that we will calculate
(2) Take a sample from the BRFSS dataset of 40,000
(3) Calculate the sample mean from this specific sample and store it in (1)
(4) Repeat (2) and (3) many many times
(5) Calculate the standard deviation of the values in (1). This is an
estimation of your standard error
```

Again what we are doing in R is creating an approximate sampling distribution and then using the standard deviation from this approximation to be our estimate for the standard error. We do this with the following code:[12]

```
> sample.means<-matrix(data=NA,nrow=1000, ncol=1) #to store the sample means
> for(i in 1:1000){
+   sample<-sample(x=brfss.df\$bmi, size=40000, replace=FALSE)
+   sample.means[i]<-mean(sample)
+ }
> sd(sample.means)
[1] 0.02359317
```

Guided Practice 4.5 In another sample of 40,000 US adults, the standard deviation of BMI is $s_y = 5.34$. Because the sample is a simple random sample and consists of less than 10% of the United States population, the observations are independent. (a) What is the standard error of the sample mean, $\bar{y} = 26.36$? (b) Would you be surprised if someone told you the average BMI of all US adults was actually 30? What about 26? [13]

Guided Practice 4.6 (a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation of the individual observations is 10, what is our estimate of the standard error when

[12] This computing experience samples without replacement to simulate experimenting in the real world. Theory states however that individual BMI values need to be independent. The 10% rule, remember, is used as a rule of thumb. By sampling without replacement within a finite population, we see that `sd(sample.means)` is not exactly the theoretical standard error but quite close.

[13] (a) Use Equation (4.4) with the sample standard deviation to compute the standard error: $SE_{\bar{y}} = 5.34/\sqrt{40000} = 0.0267$. (b) It would be surprising if the true average BMI was 30. A BMI of 30 is many standard deviations away from the sample mean of 26.36. In other words, a BMI of 30 seems implausible given that our sample mean (26.36) is far from the "true mean" using the standard error of 0.0267 to identify what is close and what is not close. Even a BMI of 26 in this situation would be surprising given that is it more than one standard deviation away from the sample mean (standard error of 0.0267).

the sample size is 100? What about when it is 400? (c) Explain how your answer to (b) mathematically justifies your intuition in part (a).[14]

4.1.4 Basic properties of point estimates

We achieved three goals in this section. First, we determined that point estimates from a sample may be used to estimate population parameters. We also determined that these point estimates are not exact: they vary from one sample to another due to sampling error. The point estimate that we observe is a single observation in the estimate's entire sampling distribution. Lastly, we quantified the uncertainty of the sample mean using what we call the standard error, mathematically represented in Equation (4.4) and through computation with R. While we could also quantify the standard error for other estimates – such as the median, standard deviation, or any other number of statistics – we will postpone these extensions until later chapters and courses.

4.2 Confidence intervals

A point estimate, we saw in Section 4.1.1 provides a single plausible value for a parameter. However, a point estimate is rarely perfect and exact; usually there is some error in the estimate as shown from the existence and calculation of standard error. Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* to estimate the true value of the parameter.

In this section and in Section 4.3, we will emphasize the special case where the point estimate is a sample mean and the parameter that we are interested in is the population mean. In Section 4.5, we generalize these methods for a variety of point estimates and population parameters that we will encounter in Chapter **??** and beyond.

4.2.1 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 in the same lake with a net. We can throw a spear where we see 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. There is likely to be some error associated with this estimate. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter within our range.

⊙ **Guided Practice 4.7** If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[15]

[14] (a) Extra observations are usually helpful in understanding the population, so a point estimate with 400 observations seems more trustworthy. (b) The standard error when the sample size is 100 is given by $SE_{100} = 10/\sqrt{100} = 1$. For 400: $SE_{400} = 10/\sqrt{400} = 0.5$. The larger sample has a smaller standard error. (c) The standard error of the sample with 400 observations is lower than that of the sample with 100 observations. The standard error describes the typical error, and since it is lower for the larger sample, this mathematically shows the estimate from the larger sample tends to be better – though it does not guarantee that every large sample will provide a better estimate than a particular small sample.

[15] 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. The more values we

⊙ **Guided Practice 4.8** Suppose we have a confidence interval that is 10 units wide and that we are 50% confident that the range encompasses the population parameter. If we had another interval that was instead 5 units wide centered at the same value as our original interval, are we now more or less confident than 50% that the range will include the population parameter? [16]

4.2.2 Confidence levels

As we have seen from the previous exercises, the size of our fishing net depends on how confident we want to be in catching a fish. Similarly, the size or width of our confidence intervals depends on how confident we want to be in estimating the true value of the parameter in question. Thus before we jump into the confidence interval itself, let's first understand what it means to be ”confident.”

Before most inference processes like creating confidence intervals and hypothesis testing, we first choose a confidence level. This confidence level is a percent that affects how wide the interval you calculate is. Confidence levels are associated with a level of uncertainty and how much you are allowing your test to commit a Type I Error which appears in Section 4.3.4. For example if you wanted to say that we are 75% confident that the population mean BMI is between two values, 75% would be our measure of uncertainty and 25% would be the Type 1 Error. Statisticians generally use a confidence level of 95% per tradition but any confidence level is allowed given varying inference goals in mind as Section 4.3.6 demonstrates.

But what does ”95% confident” truly mean? Suppose we took many samples and built a confidence interval based on each sample. Then to be 95% confident, we would see approximately 95% of those intervals would contain the actual mean, the true value, μ. In this case, if we took 100 independent samples and built 100 confidence intervals, 95 of these confidence intervals would contain the average BMI, 26.35088 and 5 of these would not.

Figure 4.8 shows this process with 25 samples, where 24 of the resulting confidence intervals contain the average BMI for the population, $\mu = 26.35088$, and one does not.

4.2.3 Confidence intervals through computation

Figure 4.8 should give you inspiration on how to achieve an estimate of a 95% confidence interval through computation. Our goal, again, is to find a range of values that hopefully contain the true population parameter. If we have the ability to resample from our population and observe many independent samples, the natural method to estimate a confidence interval is through the sampling distribution of sample means like we saw in Figure ??. We observed there that the true population parameter was extremely close to the mean of the sampling distribution. Thus a reasonable estimation for the confidence interval would be to take the middle 95% of all the sample means of the sampling distribution. Below we have pseudocode that implements this procedure (which highly resembles the pseudocode for approximating a sampling distribution from Section 4.1.3)

```
(1) Have a place to store all the sample means that we will calculate
```

include in our range, the more likely it is that this range contains the true value since the interval contains simply *more* values. However just capturing the parameter with the widest interval is not always the best when constructing a confidence interval. We can always capture the parameter with an interval going from $-\infty$ to $+\infty$ but does range does not increase our understanding of the population parameter.

[16] We are less confident than 50% that the smaller interval includes the population parameter simply because it is smaller and contains fewer values. Using a smaller net, we are less confident that we have captured the true value.

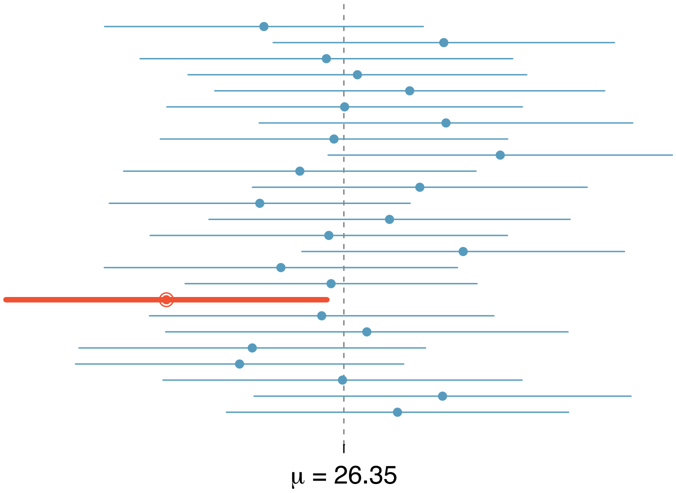


Figure 4.8: Twenty-five samples of size $n = 100$ were taken from the `BRFSS` data set. For each sample, a confidence interval was created to try to capture the average BMI for the population. Only 1 of these 25 intervals did not capture the true mean, $\mu = 26.35$.

```
(2) Take a sample from the BRFSS dataset of 40,000
(3) Calculate the sample mean from this specific sample and store it in (1)
(4) Repeat (2) and (3) many many times
(5) Use the middle 95% of values as your 95% confidence interval.
```

We can build off the method of approximation the sampling distribution to compute a 95% confidence interval. Once we have the array of 100,000 sample means (referred to as `sample.means` within the *R* code) calculated at each of the 100,000 repetitions, instead of simply plotting these values as a histogram, we can look at the cutoffs to get the middle 95% of sample means. The confidence interval itself is the BMI value c_1 such that 2.5% of the distribution is below and another BMI value c_2 such that 2.5% of the distribution values is greater than c_2. In order to find these values, recall from the distributions unit **??** that we need to use the `quantile()` function in *R*. Particularly

```
confidence.interval<-quantile(x=sample.means,c(0.025,0.975))
> confidence.interval
    2.5%    97.5%
26.30493 26.39697
```

We see that the interval (26.30493,26.39697) is an estimation for a 95% confidence interval using the sampling distribution of sample means. Therefore we can say that we are 95% confident that the true population mean BMI is between 26.30493 and 26.39697. Similarly we can also say that after calculating many many confidence intervals from many different observed samples, 95% of all the confidence intervals that we calculated will contain the true population mean.

⊙ **Guided Practice 4.9** Say we were interested in creating a 90% confidence interval and a 50% confidence interval. (a) how do you think the widths of the confidence intervals would compare? (b) How would we use the `quantile()` function to find the 90% and 50% confidence intervals from using the array `sample.means`? [17]

4.2.4 Calculating an approximate 95% confidence interval

Computing the confidence interval from resampling is straightforward and serves as a great estimate for any confidence interval for the population. However we ask ourselves, is this realistic? In general do we have the ability to resample the US population independently 100,000 times? Generally no. Most times we cannot observe 100,000 sample means of BMI values let alone 100,000 samples from the US population. More often than not, researchers only view 1 sample of 40,000 individuals and thus only 1 sample mean. How do we calculate a confidence level from only observing one sample then?

We see from the computation in Section 4.1.3 and Figure ?? that the sampling distribution is centered around the true mean and that a confidence interval is derived from the point estimate's sampling distribution. Therefore it makes sense to build the confidence interval around the point estimate that we observe as the point estimate is the most plausible value of the parameter. The standard error, which is a measure of the uncertainty associated with the point estimate, and the predetermined confidence level provide a guide for how large or small we should make the confidence interval.

With an interval at the 95% confidence level, roughly 95% of the time the estimate that you observe from resampling will be within approximately 2 standard errors[18] of the parameter, the true value. Therefore we can create an interval that is 2 standard errors from the point estimate on either side. We then can be roughly 95% **confident** that we have captured the true parameter with the confidence interval calculated by Equation 4.10 from the sample that we observe:

$$\text{point estimate } \pm\ 2 \times SE^{19} \tag{4.10}$$

Using the BMI sample from 4.1.3 with an observed sample mean of 26.35551 and standard deviation of 5.288141, we calculate the 95% confidence interval

$$\begin{gathered}
\text{point estimate } \pm\ 2 \times SE \\
26.35551 \pm 2 \times \frac{5.288141}{\sqrt{40000}} \\
26.35551 \pm 0.05288141 \\
(26.3022, 26.4079)
\end{gathered}$$

While not exact, we see that the confidence interval simulated through *R* achieves a very similar confidence interval as the one above through calculation. The difference is due to randomness within the sample since both the point estimate and the standard error vary depending on the values that we observe as well as the approximation to 2 standard errors for a 95% confidence interval.

[17] (a) We would expect the 50% confidence interval would have the smallest width. In general the more confidence we are, the larger the confidence interval width .(b) Remember we want the middle percent of observed sample means. Therefore for the 90% confidence interval, the *R* code would be `quantile(x=sample.means,c(0.05,0.95))`. For a 50% confidence interval, the code would be `quantile(x=sample.means,c(0.25,0.75))`

[18] 1.96 to be more precise if the distribution is closer to a Normal Distribution. Details coming up in Section 4.2.5

⊙ **Guided Practice 4.11** In Figure 4.8, one interval does not contain a BMI value of 26.35088. Does this imply that the true mean cannot be 26.35088? [20]

The rule where about 95% of observations are within 2 standard deviations of the mean is only approximately true. However, it holds very well for the normal distribution. As we will soon see in Section 4.4, the mean tends to be normally distributed when the sample size is sufficiently large. Similarly in the next section, we will see how to find the number of standard deviations with changing confidence levels more precisely.

⊙ **Guided Practice 4.12** The sample data suggest the average adult's age is about 46.64 years with a standard error of 0.09 years (estimated using the sample standard deviation, 17.35). What is an approximate 95% confidence interval for the average age of all of the US adults?[21]

4.2.5 A sampling distribution for the mean

In Section 4.1.3, we introduced a sampling distribution for $\bar{x}$, the average BMI value for samples of size 1,000 and 100,000. We examined this distribution earlier in Figure ??. We see with larger sample sizes, we get a more accurate depiction of the sampling distribution as stated previously. This histogram with 100,000 repetitions of random samples is shown in the left panel of Figure 4.9.

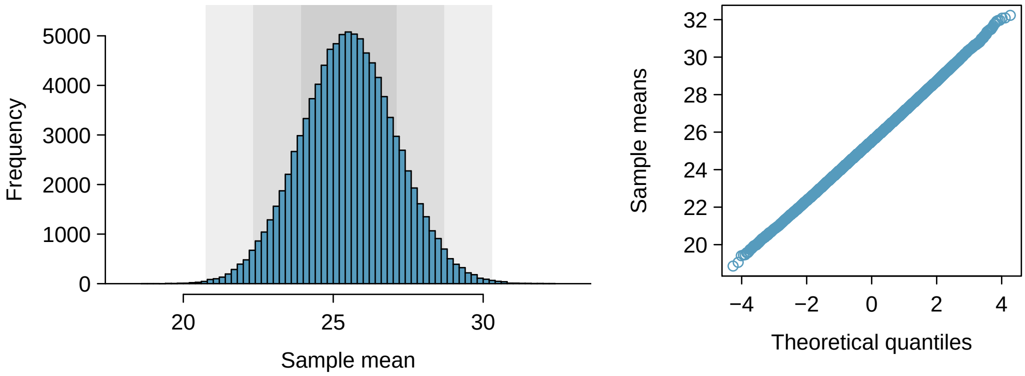


Figure 4.9: The left panel shows a histogram of the sample means for 100,000 different random samples. The right panel shows a normal probability plot of those sample means.

Does this distribution look familiar (think back to Chapter 2 of probability distributions)? Hopefully so! The distribution of sample means closely resembles the normal distribution (see Section **??**). A normal probability plot of these sample means is shown

[20] Just as some observations occur more than 2 standard deviations from the mean, some point estimates will be more than 2 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.

[21] Again apply Equation (4.10): $46.64 \pm 2 \times 0.09 \rightarrow (46.46, 46.82)$. We interpret this interval as follows: We are about 95% confident the average age of all US adults was between 46.46 and 46.82 years. Looking at the entire dataset (normally we do not have this luxury!) we see that the average age is 46.719 which is indeed within our confidence interval.

in the right panel of Figure 4.9. Because all of the points closely fall around a straight line, we can conclude the distribution of sample means is nearly normal. This result can be explained by the Central Limit Theorem[22].

Central Limit Theorem, informal description

If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model.

Why 30?

We introduce the Central Limit Theorem uses this cutoff at 30 in this text but this cutoff varies from book to book As a quick exercise both in statistical exploration but also more practice in algorithmic thinking, think about how you would visually test if 30 independent observations is a sufficient number of observations to approximate the distribution to a normal model. Let's use the `BRFSS` data to sample from like before.

Again remember we are testing if the normal model is a good approximation for the the sampling distribution of the sample mean with a sample size of 30. Creating a sampling distribution for sample sizes of $n = 5, 10, 20, 30$ and overlaying a normal approximation on the histogram is a great guide. [23]

In Figure ?? we see the sampling distributions of the sample mean for sample sizes of 5, 10, 20 and 30. The curve on top is a normal density curve with the normal distribution where μ is the mean of all the sample means and σ is the standard deviation of the sample means.

We will apply this informal version of the Central Limit Theorem for now, and discuss its details further in Section 4.4.

The choice of using 2 standard errors in Equation (4.10) was based on our general guideline that roughly 95% of the time, observations are within two standard deviations of the mean. Under the normal model, with a sufficient number of samples, we can make this more accurate by using 1.96 in place of 2.

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{4.13}$$

If a point estimate, such as $\bar{x}$, is associated with a normal model with standard error SE, then we use this more precise 1.96 to create a 95% confidence interval.

4.2.6 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% or even lower like 90%. 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.

[22] A more formal definition coming soon

[23] Use the same code for creating a sampling distribution but vary the sample size. Then use the code:
```
hist(sample.means, freq=FALSE )
curve(dnorm(x,mean=mean(sample.means), sd=sqrt(var(sample.means))), add = TRUE)
```
to add the normal curve on top of the histogram.

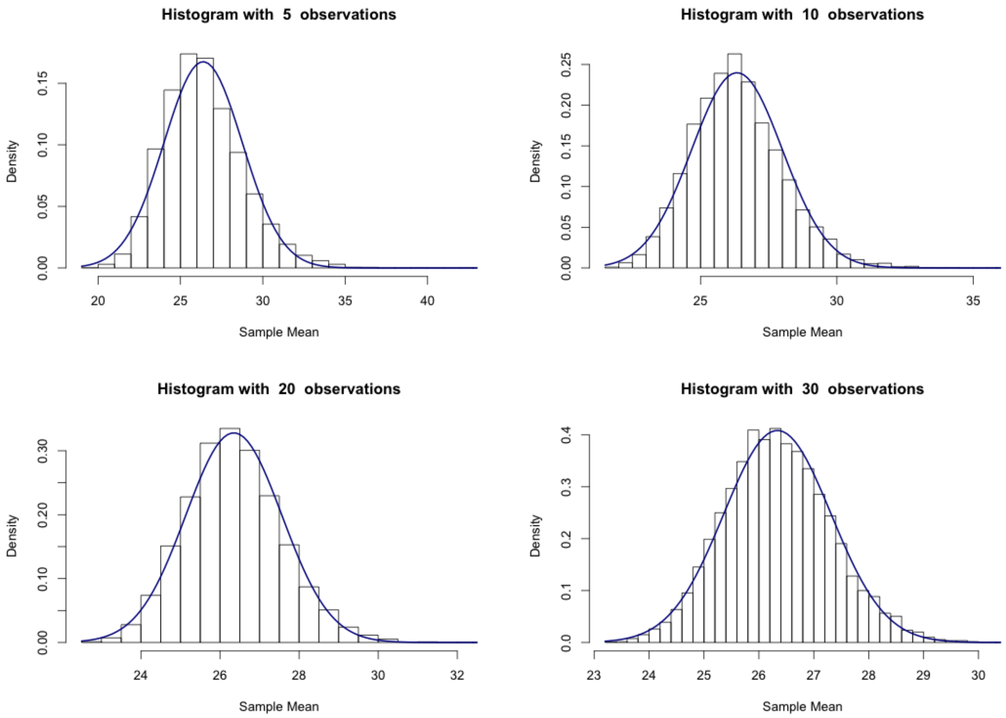


Figure 4.10: The sampling distribution of sample means with different sample sizes. With a normal density curve on top, we see that for $n = 30$, a normal model is a fitting approximation confirming the cutoff for the Central Limit Theorem.

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 where the point estimate follows a nearly normal distribution.

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{4.14}$$

There are three components to this interval: the point estimate, the "1.96", and the standard error. The $1.96 \times SE$ value affects the confidence interval width, and the point estimate affects where the confidence interval will be centered. Since we know from a normal distribution's Z-score that approximately 95% of data that is normally distributed falls within 1.96 standard deviations of the mean, $1.96 \times SE$ represents the width required to "capture that 95%" of the sampling distribution as seen in Figure ??.

⊙ **Guided Practice 4.15** If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[24]

[24] 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 ??.) 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 μ.

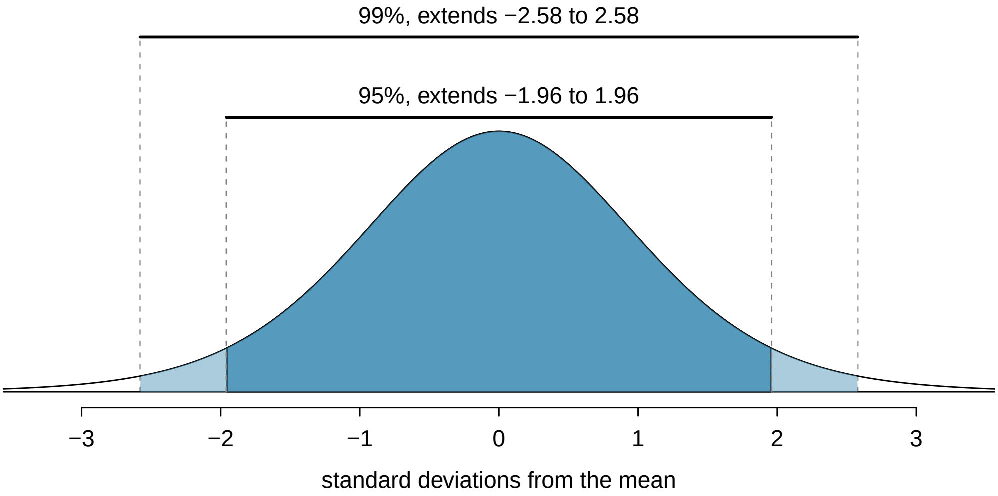


Figure 4.11: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

To 99% confident, change 1.96 in the 95% confidence interval formula to be 2.58 for a 99% confidence interval. Exercise 4.15 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean. This approach – using the Z scores in the normal model to compute confidence levels – is appropriate when $\bar{x}$ is associated with a normal distribution with mean μ and standard deviation $SE_{\bar{x}}$. Thus, the formula for a 99% confidence interval is

$$\bar{x} \ \pm\ 2.58 \times SE_{\bar{x}} \tag{4.16}$$

The normal approximation is crucial to the precision of these confidence intervals. Section 4.4 provides a more detailed discussion about when the normal model can safely be applied. When the normal model is not a good fit, we will use alternative distributions that better characterize the sampling distribution. Below however is a good checklist to determine whether or not the Central Limit Theorem can be informally applied to the distribution of sampling mean.

Conditions for $\bar{x}$ being nearly normal and SE being accurate

Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal and the estimate of SE sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The population distribution is not strongly skewed. (We check this using the sample distribution as an estimate of the population distribution.)

Additionally, the larger the sample size, the more lenient we can be with the sample's skew.

These three conditions help ensure that $\bar{x}$ is both distributed normally and representative of the target population. If the distribution of $\bar{x}$ is nearly normal, choosing a precise "1.96" or "2.58" becomes much easier for calculating confidence intervals. More importantly, however, the representativeness of the sample is imperative in our ability to infer about the target population. Randomness, independence and a large sample size safeguard against an extreme observation from skewing the conclusions from our sample. These conditions ensure the ability to accurately infer and generalize about the population of interest.

Verifying independence is often the most difficult of the conditions to check, and the way to check for independence varies from one situation to another. However, we can provide simple rules for the most common scenarios.

TIP: How to verify sample observations are independent

Observations in a simple random sample consisting of less than 10% of the population are independent.

Caution: Independence for random processes and experiments

If a sample is from a random process or experiment, it is important to verify the observations from the process or subjects in the experiment are nearly independent and maintain their independence throughout the process or experiment. Usually subjects are considered independent if they undergo random assignment in an experiment or are selected randomly for some process.

⊙ **Guided Practice 4.17** Create a 99% confidence interval for the average weight of men from the `brfss.sample` sample. The point estimate is $\bar{w} = 189.4$ and the standard error is $SE_{\bar{y}} = 0.178$. Refer to Figure 4.12 for guidance on skewness. [25]

Now that we know how to calculate a 95% and 99% confidence interval given a nearly normally distributed $\bar{x}$, we can generalize this setup to any confidence level we choose.

[25] The observations are independent (simple random sample, $<$ 10% of the population), the sample size is at least 30 ($n = 100$), and the distribution is only slightly skewed (Figure 4.12); the normal approximation and estimate of SE should be reasonable. Apply the 99% confidence interval formula: $\bar{y} \pm 2.58 \times SE_{\bar{y}} \rightarrow (188.94, 189.87)$. We are 99% confident that the average weight of all males is between 188.94 and 189.87 pounds.

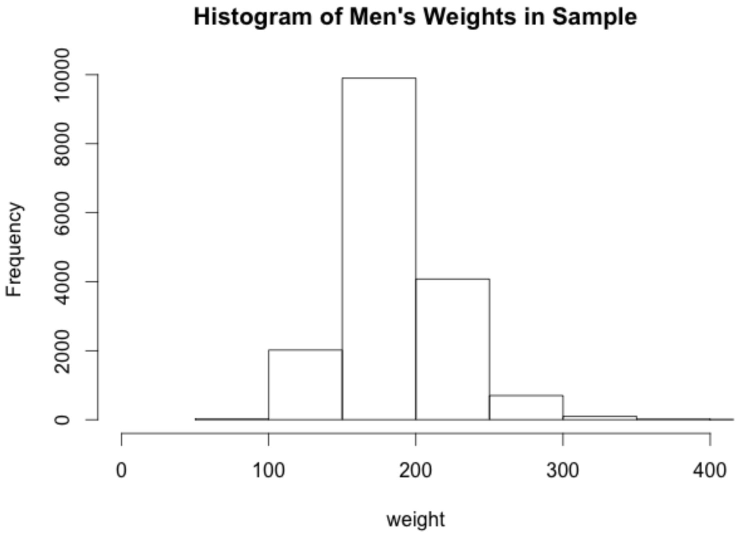


Figure 4.12: We draw a histogram of the men's weights in the `brfss.sample` and note that it is only slightly skewed. With 40,000 observations however, its skewness is more negligible because of its large sample size.

Remember while it has become tradition to use the 95% confidence level, any confidence level is allowed and vary by statistician and by goal.

Confidence interval for any confidence level (nearly normal model)

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

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

where $z^{\star}$ corresponds to the confidence level selected. The coefficient on the standard error, $z^{\star}$, is also known as the critical value. Remember that $z^{\star}$ is only used when the point estimate resembles a normal model [a]

[a] $z^{\star}$ is also used when the population standard deviation is known. However since we previously mentioned that this is rarely ever the case in practice, we have disregarded this situation completely

Margin of error

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

Figure ?? 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. We note from Figure ?? that the $z^\star$ value comes from a $\mathcal{N}(0,1)$. Therefore we can either use R or a Z-table [26] (**FOUND IN THE BACK OF THE BOOK HERE**) to find the critical value associated with some confidence level. In R, we use the `qnorm()` function. `qnorm()` takes in a probability p and outputs the quantile value z such that $P(Z \leq z) = p$. For a 95% confidence interval, $p = 0.025$ since we are looking for the *middle* 95%. Therefore in R

```
> qnorm(0.025)
[1] -1.959964
```

and we show that $z^\star = 1.96$ is the critical value for 95%.

⊙ **Guided Practice 4.18** What is the critical value associated with (a) 90%, (b) 75% and (c) 50%? [27]

⊙ **Guided Practice 4.19** Use the data in Exercise 4.17 to create a 90% confidence interval for the average weight of men in the United States.[28]

4.2.7 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...

Looking back to **Section 4.2.2**, this means that if we took a random sample from our population 100 times and calculated a confidence interval around our point estimate each time, 95 confidence intervals would contain the true population parameter.

It is interesting to note, however, that researchers in practice would almost never be able to resample 100 times and generate 100 confidence intervals. The meaning of being "95% confident" has traditionally been one grounded in theory and less in practice. "Confidence" relates more to the reliability of the process of creating such a range and less so in the probability that the value is within the range.

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, a percent of all the data or just the sampled data. A confidence interval also says nothing about capturing point estimates since the confidence interval is always centered at the observed point estimate. Confidence intervals only attempt to capture population parameters as statistical inference's goal is to infer about such population parameters.

[26] also known as a Normal table

[27] Remember we want the *middle* and for any given confidence level C, we type into R, `qnorm(`$0.5{\cdot}(1{-}C)$`)`. Therefore (a) `qnorm(0.05)= -1.644854` so $z^\star = 1.65$ for a 90% confidence level (b) 1.15 (c) 0.67

[28] We first 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 $\bar{y} \pm 1.65 \times SE_{\bar{y}} \rightarrow (189.11, 189.69)$. (We had already verified conditions for normality and the standard error in the previous exercise.) That is, we are 90% confident the average weight of males is between 189.11 and 189.69 pounds. Also note that because we are at a 90% confidence level, our confidence interval width is smaller than in Exercise 4.17.

Some incorrect interpretations of a 95% confidence interval include:

95% of the observed data is between ...
95% of the population distribution is contained in the confidence interval.

Remember, a confidence interval is not a range of plausible values for the sample mean, though it may be understood as an estimate of plausible values for the population parameter. A particular confidence interval of 95% calculated from an experiment does not mean that there is a 95% probability of a sample mean from a repeat of the experiment falling within this interval.[13]

While the differences in correct and incorrect interpretations are extremely nuanced, the goal of this book is to provide the tools and mechanisms of calculating and computing a confidence interval from data and less so about the wording which, in practice, has become almost meaningless and obsolete.

4.2.8 Nearly normal population with known SD (special topic)

In rare circumstances we know important characteristics of a population. For instance, we might already know a population is nearly normal and we may also know its parameter values. Even so, we may still like to study characteristics of a random sample from the population. Consider the conditions required for modeling a sample mean using the normal distribution:

(1) The observations are independent.

(2) The sample size n is at least 30.

(3) The data distribution is not strongly skewed.

These conditions are required so we can adequately estimate the standard deviation of the population from our sample and so we can ensure the distribution of sample means is nearly normal. However, if the population is known to be nearly normal, we know that the sample mean is always nearly normal (this is a special case of the Central Limit Theorem). If the standard deviation for the population is also known, then conditions (2) and (3) are not necessary for those data.

We would like to heavily emphasize however that while, in practice, the population mean is more likely to be known, the population standard deviation is rarely know. While a known population standard deviation will rarely occur in practice, the Central Limit Theorem allows us to describe the distribution of the sampling distribution more specifically.

● **Example 4.20** The heights of male seniors in high school closely follow a normal distribution $N(\mu = 70.43, \sigma = 2.73)$, where the units are inches.[29] If we randomly sampled the heights of five male seniors, what distribution should the sample mean follow?

The population is nearly normal, the population standard deviation is known, and the heights represent a random sample from a much larger population, satisfying the independence condition. Therefore the sample mean of the heights will follow a nearly normal distribution with mean $\mu = 70.43$ inches and standard error $SE = \sigma/\sqrt{n} = 2.73/\sqrt{5} = 1.22$ inches.

[29] These values were computed using the USDA Food Commodity Intake Database.

Alternative conditions for applying the normal distribution to model the sample mean

If the population of cases is known to be nearly normal and the population standard deviation σ is known, then the sample mean $\bar{x}$ will follow a nearly normal distribution $N(\mu, \sigma/\sqrt{n})$ if the sampled observations are also independent.

Sometimes the mean changes over time but the standard deviation remains the same. In such cases, a sample mean of small but nearly normal observations paired with a known standard deviation can be used to produce a confidence interval for the current population mean using the normal distribution.

TIP: Relaxing the nearly normal condition

As the sample size becomes larger, it is reasonable to *slowly* relax the nearly normal assumption on the data when dealing with small samples. By the time the sample size reaches 30, the data must show strong skew for us to be concerned about the normality of the sampling distribution.

4.3 Hypothesis testing

Is the average US adult satisfied with his or her weight? We consider this question in the context of the BRFSS dataset comparing US adults' current weight and their desired weight (we will call this ”weight difference”). While media pressures women to maintain a slim figure, the same media urges men to work out more and become stronger and more fit. These opposing viewpoints and many others all are components that influence satisfaction with weight and the desire to lose or gain weight.

In addition to considering weight in this section, we consider a topic near and dear to most students: sleep. A recent study found that college students average about 7 hours of sleep per night.[30] However, researchers at a rural college are interested in showing that their students sleep longer than seven hours on average. We investigate this topic in Section 4.3.2.

Many questions, given the correct data, can be answered through Hypothesis Testing. **Hypothesis testing** is a method in statistics that evaluates whether or not a population parameter has a hypothesized value with an associated probability of error. It is, most obviously, determining the probability that a given hypothesis is true or not.

Hypotheses are often simple questions that have a yes or no answer. Consider some hypotheses below:

Is the mean body temperature really 98.6F?
Has consumption of soda changed across the US overtime?
Do MCAT classes improve MCAT scores?

The hypothesis testing process consists of generally 5 steps. Going through the **Hypothesis testing framework** allows for statisticians to answer these yes/no questions with a certain degree of confidence after observing a related sample. We begin by testing a

[30] http://theloquitur.com/?p=1161

hypothesis about a population mean from observing one sample. Remember, we can do hypothesis testing on any population parameter. It can be the population mean, population standard deviation or even the population IQR if desired.

4.3.1 Hypothesis testing framework

The average weight difference that adults want to experience that we observe from our sample of the `brfss.sample` data is 15.01 lbs. We want to determine if this sample provides enough evidence that adults are satisfied with their weight versus the alternative – that they are not.[31] We use desired weight difference as a proxy for weight satisfaction and simplify this question into two **hypotheses**

H_0: US adults are satisfied with their current weight. The average desired weight difference for US adults is 0 lbs.

H_A: The average adult's desired weight difference is not 0 lbs i.e. Average adults are not satisfied with their current weights and would like to change.

Step 1: Formulating Hypotheses

The first step within the hypothesis testing framework is setting up the hypotheses. As shown above, we generally have two hypotheses, a null and an alternative.

We call H_0 the null hypothesis and H_A the alternative hypothesis.

H_0
null hypothesis

H_A
alternative hypothesis

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 parameter values.

The null hypothesis often represents a skeptical position. The null hypothesis is generally denoted as "no difference" or what one would observe if there is no change. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change. If the null hypothesis is true, any difference between the observed sample is due only to chance variation.

TIP: Hypothesis testing framework

The logic of hypothesis testing is that we will not reject the null hypothesis (H_0), unless the evidence in favor of the alternative hypothesis (H_A) is so strong that we must reject H_0 in favor of H_A.

The first step within 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. We believe our null hypothesis H_0. However, if there is sufficient evidence that we observe that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative.

[31] While we could answer this question by examining the entire population data (`BRFSS`), we only consider the sample data (`brfss.sample`), which is more realistic since statisticians rarely have access to population data.

⊙ **Guided Practice 4.21** A new study would like to be published in a scientific journal. The board that determines the validity of the study considers two possible claims about this study: either the study is valid or pseudoscience. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative? [32]

Those scientists who sit on the board of publication journals look at the study, previous literature and other evidence to see whether it convincingly supports that the science is valid. Even if these scientists leave unconvinced that the study is publishable, this does not mean that these board members believe the study is complete fabrication. 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 true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

TIP: Double negatives can sometimes be used in statistics
In many statistical explanations, we use double negatives. For instance, we might say that the null hypothesis is *not implausible* or we *failed to reject* the null hypothesis. Double negatives are used to communicate that while we are not rejecting a position, we are also not saying it is correct.

In the example with the `BRFSS` data, the null hypothesis represents no change in desired weight difference. The alternative hypothesis represents something new or more interesting: there was a difference, either a desire to gain or lose weight on average. These hypotheses can be described in mathematical notation using μ_{wd} as the average weight difference for US adults.

$$H_0 : \mu_{wd} = 0 \qquad\qquad H_A : \mu_{wd} \neq 0$$

where 0 represents a desired weight difference of 0 lbs or that these US adults on average do not care to change their weight. Using this mathematical notation, the hypotheses can now be evaluated using statistical tools. We call 0 the **null value** since it represents the value of the parameter if the null hypothesis is true. We will use the `brfss.sample` data set to evaluate the hypothesis test.

Note it is important to remember that we are not testing whether or not the average weight difference observed from the `brfss.sample` is 0 or not. We don't need to test that since we have observed all of `brfss.sample` and can simply calculate it. Rather we are testing the *population parameter* or the true average value of all US adults' weight differences is 0 or not.

[32] The board considers whether the study's evidence, results and reproducibility is so convincing (strong) that there the study must be valid. In this case the board rejects the null hypothesis (the study is pseudoscience) and concludes that the study is valid an should be published (alternative hypothesis).

TIP: Null and Alternative Hypothesis Setup

The null hypothesis is generally written as $H_0 : \mu = \mu_0$ where μ is the population mean and μ_0 is the hypothesized value that we believe to be true.
The alternative hypothesis, on the other hand, can be many things.
If we have no prior belief to influence our alternative hypothesis and the researchers are interested in showing any difference –an increase or decrease– then the safest one would be $\mu \neq \mu_0$ or a two-sided alternative. If we have a prior belief of how μ and μ_0 compare or are interested in only showing an increase or decrease, but not both, we can do a one-sided alternative, $\mu \geq \mu_0$ or $\mu \leq \mu_0$. We will go into more detail on one-side versus two sided in Section 4.3.2.

Step 2: Specifying a Significance Level α

Once we have completed Step 1 and have a null and alternative hypothesis, we need to specify a **significance level**. The significance level α is the acceptable error probability of the test. In this case, the error probability is the probability of concluding the alternative hypothesis is true when it is not true. This error is called a Type I error, and α is the probability of a Type I error. We will go into more detail on error types in Section 4.3.4.

Typically, α is taken to be 0.05, 0.01, or some other small value. α plays the same role as the error probability in confidence intervals, and is a measure of uncertainty. If $\alpha = 0.05$, we are testing at a 95% confidence level for our hypothesis tests. We will see a clearer connection between hypothesis testing and confidence intervals in Section 4.3.3.

Step 3: Calculating the Test Statistic

The third step is calculating a test statistic from the data we observe. This statistic will be the value that the conclusions will be based on and measures the difference between the observed data and what is expected if the null hypothesis is true. This test statistic answers the question: "How many standard deviations from the hypothesized value is the observed sample value?" Thinking back to **THIS SECTION ??** by standardizing a normal, the test statistic follows a similar construction. When testing hypotheses about a mean, the test statistic for the population mean from one sample will always be

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the number of observations in the sample. *Note:* In general we see that test statistics follow $\frac{\text{observed}-\text{hypothesized}}{\text{standarderror}}$ to see how many standard deviations the observed value is from the hypothesized value. This T-statistic follows a t-distribution [33] and will have $n-1$ degrees of freedom.

[33] from **Chapter 3**

Test statistic

A *test statistic* is a special summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. The test-statistic is a particular data summary that summarizes how many standard deviations from the hypothesized null value is the observed sample value. In general the T-statistic follows a t-distribution with $n - 1$ degrees of freedom. [a]

[a] When a point estimate is nearly normal, we use the Z score of the point estimate as the test statistic. In later chapters we encounter situations where other test statistics are helpful.

Step 4: Calculating the p-value

Once we calculate a test statistic from the observed data, we know how many standard deviations our observation is from the hypothesized value if the null hypothesis were true. Now we need to tie this T-statistic value to a probability of such an observation happening. We do this through the **p-value**. Assuming the null hypothesis is true, the p-value is the probability of observing our sample or a more extreme sample. Formally the p-value is a conditional probability.

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 is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

How do we get this probability? Section 4.3.2 will go into more detail from using R, Z and t- tables.

Step 5: Making your conclusion

The final step within the hypothesis testing framework is to make a conclusion from the p-value we calculated in Step 4. Using the definition of p-value, if we observe something extreme, the probability associated with our observation will be small. Thus if our observation is rare, the T-statistic and p-value provide evidence that the hypothesized value is unlikely. Therefore if our p-value is low, we should reject our null hypothesis, and the smaller the p-value, the stronger the evidence we have against the null hypothesis.

How small is small? This is where Step 2 and our significance level comes in. If the p-value is small or smaller than the pre-specified α level (usually 0.01 or 0.05), we reject the null hypothesis and say the result that we observe is statistically significant at the α level.

If the p-value is α or greater, we simply do not have enough evidence to reject the null hypothesis. The subtle but important point is that not rejecting H_0 is not equivalent to accepting H_0 (refer back to Example 4.21). In practice, however, not rejecting H_0 is equivalent to accepting H_0 when making decisions and acting on conclusions. Most importantly, it is key that students state the conclusion in the context of the original problem, using the language and units of that problem. Most students forget this but is absolutely necessary in both theory and practice.

4.3.2 Calculating p-values

Calculating p-values can be the most difficult part of the hypothesis testing framework. The p-value depends on many moving parts, including the sample mean, the sample size and the alternative hypothesis but always remember that the p-value is the probability of observing data as extreme or more if we assume the null hypothesis is true with the data at least as favorable to the alternative hypothesis. If the p-value is small, then our sample indicates that we just observed something rare, so rare that we should probably reject the null hypothesis as true. Figure 4.13 shows the distribution of the sample mean where the p-value is the shaded area for a one sided alternative $\mu > \mu_0$.

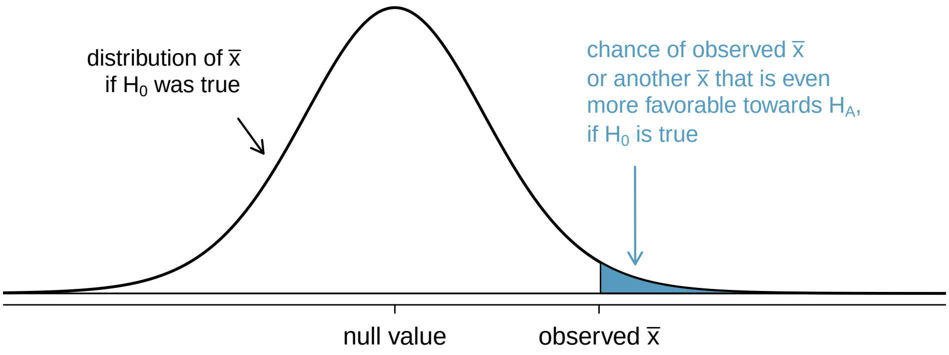


Figure 4.13: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of observing the observed $\bar{x}$ or an $\bar{x}$ even more extreme and thus favorable to follow H_A under this distribution.

If the alternative is one sided and has the form $\mu > \mu_0$, then the p-value would be represented by the upper tail (Figure 4.13). If the alternative is one sided but has the form $\mu < \mu_0$, then the p-value would be the shaded area in the lower tail. In a two-sided test, *we shade two tails* since evidence in either direction is favorable to H_A (Figure 4.18).

Now that we know what the p-value represents, how do we actually get this shaded area to be a number? Here is where the T-statistic comes into play. Before we get to the

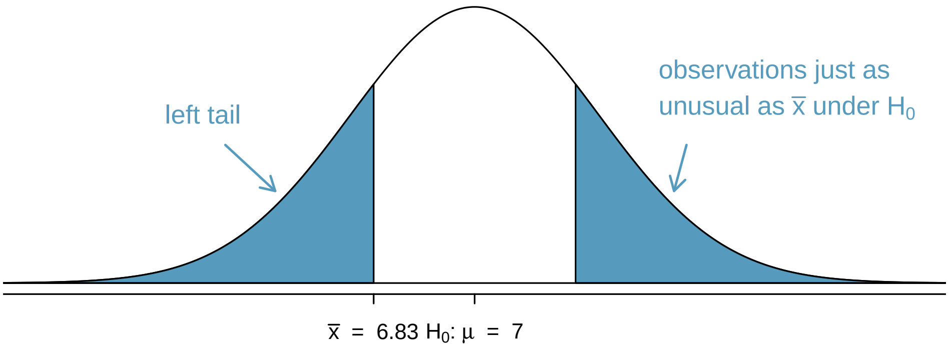


Figure 4.14: H_A is two-sided, so *both* tails must be counted for the p-value.

nitty gritty, let's look back to the BRFSS data.

Recall that the researchers for the BRFSS data are interested if US adults are satisfied with their current weight. They believe that the desired weight difference is a good proxy to measure satisfaction and have the following null and alternative hypotheses where μ_{wd} denotes the average desired weight difference in the US:

H_0: $\mu_{wd} = 0$

H_A: $\mu_{wd} \neq 0$

Instead of 40,000 within our sample, let's say that we observed a sample of 100 people and calculated a sample mean of weight differences of 0.5 pounds and standard deviation of 5 pounds. Given this information we can first calculate a T-statistic[34].

$$t = \frac{0.5 - 0}{25/\sqrt{100}} = 0.2$$

The T-statistic can be thought of as a Z-score (standard score) that indicates how many standard deviations the observed sample mean is from the null value. This standardization becomes a great way to unify all the moving parts in order to calculate the p-value.

With the T-statistic and the alternative hypothesis, we calculate the p-value from either a t-distribution or a normal distribution. The sample size determines which distribution to model our point estimate from, either a t-distribution or a normal distribution. If $n \geq 30$ **from this part**, the sample mean can be thought of coming from a normal distribution. If $n < 30$, model the sampling distribution from a t-distribution.

With $\alpha = 0.05$, students can either use a table to calculate the p-value or use R. First assuming a t-distribution, use the t-table given **on some page** to find the row with the correct degrees of freedom (in a one sample test, $df = n - 1$). Students then should look across that row to find the T-statistic value that they calculated. Note that the table won't have every single value listed but once they find the approximate T-statistic, look at the top of the column to get p-value using either a one sided or two sided (one tail or two tail) alternative. The normal table (Z-table) is very similar but be wary of that the Z-table only lists the areas left of the Z-score. This simply means that these probabilities coincide with a one-sided alternative. However because the normal distribution is symmetric, finding the p-value for a two sided alternative is just those values from the table times two!

If available, R is also a handy tool. Use the `pt()` or the `pnorm()` function to calculate the area left of the T-statistic. Students then can take the value and subtract from one or multiply by two depending on the alternative hypothesis. If students have the ability to use R, the $n \geq 30$ threshold can be loosened since modeling after the t-distribution becomes easier and more accurate compared to the tables. However we note again that once $n \geq 30$, both distributions become almost equal.

We use a normal table for calculating the p-value for our sample from the BRFSS data because $n = 100$ in our sample. A score of 0.2 corresponds to a shaded area of 0.579 to the left. Therefore in the tail we have

$$\begin{aligned} p &= Pr(T \leq -0.2) + Pr(T \geq 0.2) \\ &= Pr(|T| \geq 0.2) \\ &= 2Pr(T \geq 0.2) \\ &= 2 \cdot (1 - 0.579) \\ &= 0.842 \end{aligned}$$

[34] calculating the T-statistic using actual data is an exercise in the book

Using R, we use both `pnorm()` and `pt()` to check.

```
> 2*(1-pnorm(0.2))
[1] 0.8414806
> 2*(1-pt(0.2, 99))
[1] 0.8418908
```

We see that both output p-values that are extremely similar and agree with the p-value from the normal table as well. Now that we calculated the p-value we can conclude that this p-value $> \alpha = 0.05$ so we cannot reject H_0. To put it into context: from observing a sample mean of 0.5 for weight differences, we observed a p-value of 0.84 and cannot conclude that weight difference is nonzero. From our sample it appears that US adults are satisfied with their weight.

Guided Practice 4.22 If the null hypothesis is true, how often should the p-value be less than 0.05?[35]

TIP: Concluding on Critical Values

Conclusions are made from the p-value but if $\alpha = 0.05$ or some other common value, we can take a quick shortcut using the critical value. We learned the critical value as the coefficient on the standard error to calculate the confidence interval. However the critical value is also the point on the test distribution that can be compared to the T-statistic in hypothesis testing. Since we know that the critical value is associated with some confidence level, this critical value is also associated with α. If the absolute value of the T-statistic is greater than the critical value (more extreme), the p-value is less than α and you can reject the null hypothesis.

Caution: Critical value $\neq$ test statistic

Many times students get confused between the critical value and the test statistic. The critical value is associated with some α and does not change. For a specific α there is only one critical value. The T-statistic can change depending on the sample that you observed. Students are comparing their T-statistic to the critical value using the critical value as a benchmark.

Guided Practice 4.23 A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. Researchers at a rural school are interested in showing that students at their school sleep longer than seven hours on average, and they would like to demonstrate this using a sample of students. What would be an appropriate skeptical position for this research?[36]

We can set up the null hypothesis for this test as a skeptical perspective: the students at this school average 7 hours of sleep per night. The alternative hypothesis takes a new form reflecting the interests of the research: the students average more than 7 hours of sleep. We can write these hypotheses as

$$H_0 : \mu = 7 \qquad\qquad H_A : \mu \geq 7$$

[35] About 5% of the time. If the null hypothesis is true, then the data only has a 5% chance of being in the 5% of data most favorable to H_A.

[36] A skeptic would have no reason to believe that sleep patterns at this school are different than the sleep patterns at another school.

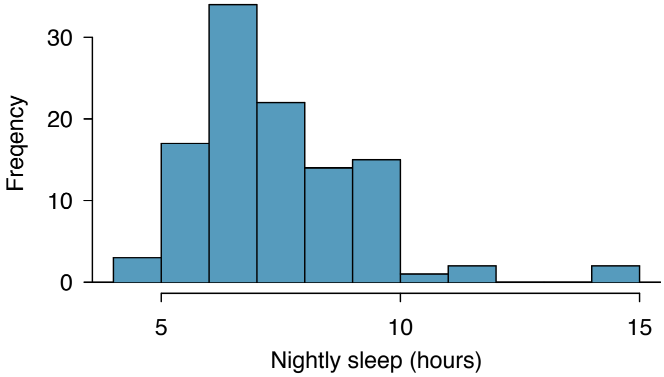


Figure 4.15: Distribution of a night of sleep for 110 college students. These data are moderately skewed.

Using $\mu \geq 7$ as the alternative is an example of a **one-sided** hypothesis test mentioned previously. In this investigation, there is no apparent interest in learning whether the mean is less than 7 hours.[37] Earlier we encountered a **two-sided** hypothesis where we looked for any clear difference, greater than or less than the null value.

Always use a two-sided test unless it was made clear prior to data collection that the test should be one-sided. Switching a two-sided test to a one-sided test after observing the data is dangerous because it can inflate the chance of an incorrect conclusion.

TIP: One-sided and two-sided tests

If the researchers are only interested in showing an increase or a decrease, but not both, use a one-sided test. If the researchers would be interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

TIP: Always write the null hypothesis as an equality

We will find it most useful if we always list the null hypothesis as an equality (e.g. $\mu = 7$) while the alternative always uses an inequality (e.g. $\mu \neq 7$, $\mu \geq 7$, or $\mu \leq 7$).

The researchers at the rural school conducted a simple random sample of $n = 110$ students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours. A histogram of the sample is shown in Figure 4.15.

Before we can use a normal model for the sample mean or compute the standard error of the sample mean, we must verify conditions. (1) Because this is a simple random sample from less than 10% of the student body, the observations are independent. (2) The sample size in the sleep study is sufficiently large since it is greater than 30. (3) The data show moderate skew in Figure 4.15 and the presence of a couple of outliers. This skew and the

[37] This is entirely based on the interests of the researchers. Had they been only interested in the opposite case – showing that their students were actually averaging fewer than seven hours of sleep but not interested in showing more than 7 hours – then our setup would have set the alternative as $\mu \leq 7$.

outliers (which are not too extreme) are acceptable for a sample size of $n = 110$. With these conditions verified, the normal model can be safely applied to $\bar{x}$ and the estimated standard error will be very accurate.

⊙ **Guided Practice 4.24** What is the standard deviation associated with $\bar{x}$? That is, estimate the standard error of $\bar{x}$.[38]

The hypothesis test will be evaluated using a significance level of $\alpha = 0.05$. We want to consider the data under the scenario that the null hypothesis is true. In this case, the sample mean is from a distribution that is nearly normal and has mean 7 and standard deviation of about 0.17. Such a distribution is shown in Figure 4.16.

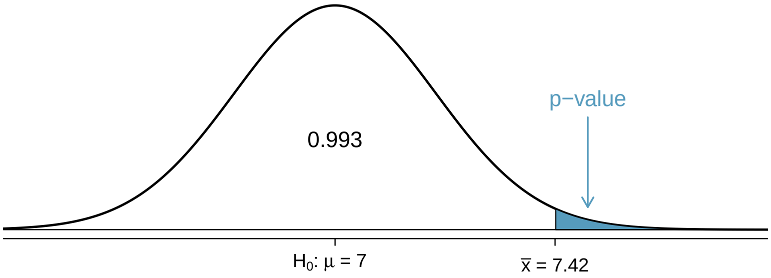


Figure 4.16: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

Remember the shaded tail in Figure 4.16 is the p-value and so we shade all means larger than our sample mean, $\bar{x} = 7.42$, because they are more favorable to the alternative hypothesis than the observed mean. We compute the p-value by first computing the T-statistic for the sample mean, $\bar{x} = 7.42$:

$$T = \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

Using the normal probability table, the lower unshaded area is found to be 0.993. Thus the shaded area is $1 - 0.993 = 0.007$. Using R we have

```
 > 1-pnorm(2.47)
[1] 0.006755653
```

If the null hypothesis is true, the probability of observing such a large sample mean for a sample of 110 students is only 0.007. That is, if the null hypothesis is true, we would not often see such a large mean.

We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level (p-value $= 0.007 < 0.05 = \alpha$), we reject the null hypothesis.[39] What we observed is so unusual with respect to the null hypothesis that it casts serious doubt on H_0 and provides strong evidence favoring H_A.

[38] The standard error can be estimated from the sample standard deviation and the sample size: $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} = 0.17$.

[39] Using critical values instead, we know that for $\alpha = 0.05$ and a one sided alternative, the critical value is 1.65. Since our T-statistic is greater than 1.65, we know to reject H_0 without calculating the actual p-value

p-value as a tool in hypothesis testing

The p-value quantifies how strongly the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

TIP: It is useful to first draw a picture to find the p-value

It is useful to draw a picture of the distribution of $\bar{x}$ as though H_0 was true (i.e. μ equals the null value), and shade the region (or regions) of sample means that are at least as favorable to the alternative hypothesis. These shaded regions represent the p-value.

⊙ **Guided Practice 4.25** Suppose we had used a significance level of 0.01 in the sleep study. Would the evidence have been strong enough to reject the null hypothesis? (The p-value was 0.007.) What if the significance level was $\alpha = 0.001$? [40]

4.3.3 Testing hypotheses using confidence intervals

While confidence intervals may appear separate from hypothesis testing, these two concepts arrive as the same conclusions. Consider a sample of 100 people from the `BRFSS` data to test if the average age of adults is 36.8 years [41]. The hypothesis setup would be $H_0 : \mu_{\text{age}} = 36.8$ and $H_A : \mu_{\text{age}} \neq 36.8$. We learned in Section 4.1 that there is fluctuation from one sample to another, and it is very unlikely that the sample mean will be exactly equal to our parameter; we should not expect $\bar{x}_{\text{ages}}$ to exactly equal μ_{ages} and the difference could be due to *sampling variation*, i.e. the variability associated with the point estimate when we take a random sample.

In Section 4.2, confidence intervals were introduced as a way to find a range of plausible values for the population mean. From `BRFSS`, the sample has a mean of 46.48 and a standard deviation of 16.83. Therefore the 95% confidence interval is

$$46.48 \pm 1.96 \cdot \frac{16.83}{\sqrt{100}} = (43.1796, 49.7804)$$

Because 36.8 years does not fall in the range of plausible values, we can say the null hypothesis is implausible. That is, we failed to reject the null hypothesis, H_0.

⊙ **Guided Practice 4.26** An investigator is studying the results of standardized IQ tests in adolescents who suffered from severe asthma during childhood. She claims that those who had childhood asthma perform worse. For the standardized test she will use, the population mean score is 100. What are the null and alternative hypotheses to test whether this claim is accurate? [42]

● **Example 4.27** In her sample of 100 children, she found a sample mean $\bar{x} = 96.7$ and standard deviation $s = 10$. Construct a 95% confidence interval for the population mean and evaluate the hypotheses of Exercise 4.26.

[40] We reject the null hypothesis whenever *p-value* $< \alpha$. Thus, we would still reject the null hypothesis if $\alpha = 0.01$ but not if the significance level had been $\alpha = 0.001$.

[41] as calculated by the US Census in 2009

[42] H_0: The average score is 100, $\mu = 100$. H_A: The average score is lower, $\mu \leq 100$.

$$SE = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

The normal model may be applied to the sample mean because the conditions are met: The data are a simple random sample and we assume that there are more than 1,000 adolescents who have suffered from asthma. The observations are independent and the sample size is also sufficiently large (n=100). We don't know about existing outliers but the sample size mitigates potential effects of outliers. This ensures a 95% confidence interval may be accurately constructed:

$$\bar{x} \ \pm \ z^{\star} SE \quad \rightarrow \quad 96.7 \ \pm \ 1.96 \times 1 \quad \rightarrow \quad (94.74, 98.66)$$

Because the null value 100 is not in the confidence interval, a true mean of 100 is implausible and we reject the null hypothesis. The data provide statistically significant evidence that adolescents who suffered from severe asthma during childhood do perform worse on standardized IQ tests.

4.3.4 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, we can make a wrong decision in statistical hypothesis tests. However, the difference is that we have the tools necessary to quantify how often we make such errors.

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 4.17.

		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 4.17: Four different scenarios for hypothesis tests.

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

⊙ **Guided Practice 4.28** 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 4.17 may be useful.[43]

⊙ **Guided Practice 4.29** How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?[44]

[43] 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).

[44] 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.

⊙ **Guided Practice 4.30** How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[45]

⊙ **Guided Practice 4.31** Consider a person getting tested for HIV. What does a Type 1 and Type 2 Error represent in this context? [46]

Exercises 4.28-4.30 provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject H_0 unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time. This corresponds to a **significance level** of 0.05 which is the same significance level from hypothesis testing and confidence intervals. We often write the significance level using α where $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 4.3.6.

α
significance
level of a
hypothesis test

If we use a 95% confidence interval to test a hypothesis where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

4.3.5 Two-sided versus One-sided hypothesis testing: Dos and Don'ts

Determining an alternative hypothesis can get tricky, and the choice between a one-sided and two sided test can be controversial. In this book, the examples and exercises will be obvious enough to decide a correct alternative hypothesis. In practice with real world data, however, can be less straightforward. If the sidedness is uncertain, many scientists opt to use a two-sided alternative because it is more *conservative*. What does conservative in this context mean? Let's first consider the differences.

It is never okay to change two-sided tests to one-sided tests after observing the data. In this example we explore the consequences of ignoring this advice. Using $\alpha = 0.05$, we show that freely switching from two-sided tests to one-sided tests will cause us to make twice as many Type 1 Errors as intended. [47]

⊙ **Guided Practice 4.32** Earlier we talked about a research group investigating whether the students at their school slept longer than 7 hours each night. Let's consider a second group of researchers who want to evaluate whether the students at their college differ from the norm of 7 hours. Write the null and alternative hypotheses for this investigation.[48]

[45] 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.

[46] Type 1 Error is if this person does not have HIV but was tested positive for HIV. Type 2 Error would be failing to detect HIV when the patient actually has HIV.

[47] hence to be conservative and safe, we opt to minimize the Type 1 Errors and use the two sided alternative

[48] Because the researchers are interested in any difference, they should use a two-sided setup: $H_0 : \mu = 7$, $H_A : \mu \neq 7$.

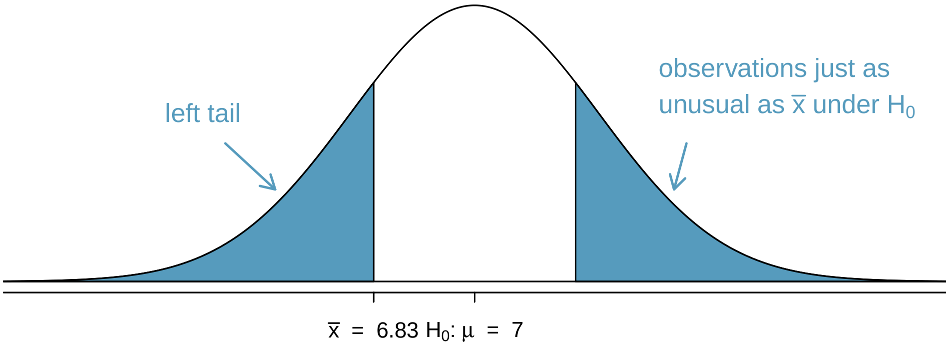


Figure 4.18: H_A is two-sided, so *both* tails must be counted for the p-value.

● **Example 4.33** The second college randomly samples 72 students and finds a mean of $\bar{x} = 6.83$ hours and a standard deviation of $s = 1.8$ hours. Does this provide strong evidence against H_0 in Exercise 4.32? Use a significance level of $\alpha = 0.05$.

First, we must verify assumptions. (1) A simple random sample of less than 10% of the student body means the observations are independent. (2) The sample size is 72, which is greater than 30. (3) Based on the earlier distribution and what we already know about college student sleep habits, the distribution is probably not strongly skewed.

Next we can compute the standard error ($SE_{\bar{x}} = \frac{s}{\sqrt{n}} = 0.21$) of the estimate and create a picture to represent the p-value, shown in Figure 4.18. Both tails are shaded. An estimate of 7.17 $(6.83 + 1.65 cdot 0.21)$ or more provides at least as strong of evidence against the null hypothesis and in favor of the alternative as the observed estimate, $\bar{x} = 6.83$.

We can calculate the tail areas by first finding the lower tail corresponding to $\bar{x}$:

$$T = \frac{6.83 - 7.00}{0.21} = -0.81 \quad \overset{table}{\rightarrow} \quad \text{left tail} = 0.2090$$

Because the normal model is symmetric, the right tail will have the same area as the left tail. The p-value is found as the sum of the two shaded tails:

$$\text{p-value} = \text{left tail} + \text{right tail} = 2 \times (\text{left tail}) = 0.4180$$

This p-value is relatively large (larger than $\alpha = 0.05$), so we should not reject H_0. That is, if H_0 is true, it would not be very unusual to see a sample mean this far from 7 hours simply due to sampling variation. Thus, we do not have sufficient evidence to conclude that the mean is different than 7 hours.

● **Example 4.34** Let's consider two cases: (1) The sample mean was larger than the null value and (2) the sample mean as smaller than the null value.

Suppose the sample mean was larger than the null value, μ_0 (e.g. μ_0 would represent 7 if H_0: $\mu = 7$). Then if we can flip to a one-sided test instead of a two-sided test, we would use H_A: $\mu > \mu_0$. Now if we obtain any observation with a T-statistic greater than 1.65, we would reject H_0. If the null hypothesis is true, we incorrectly reject the

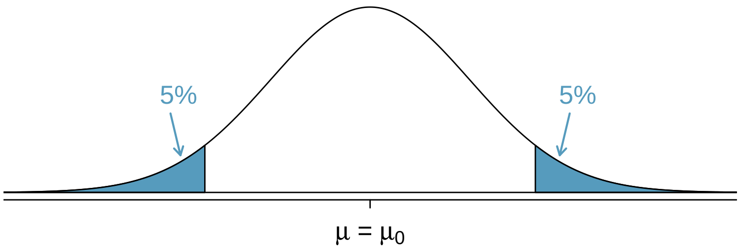


Figure 4.19: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example 4.3.5 when $\alpha = 0.05$.

null hypothesis about 5% of the time when the sample mean is above the null value, as shown in Figure 4.19.

Suppose the sample mean was smaller than the null value. Then if we change to a one-sided test, we would use H_A: $\mu < \mu_0$. If $\bar{x}$ had a T-statistic smaller than -1.65, we would reject H_0. If the null hypothesis is true, then we would observe such a case 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: One-sided hypotheses are allowed only *before* seeing data

After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Remember, the direction of a one-sided test must be made a priori, not after peeking at the data since the results could be statistically significant with a one-sided test, but not significant with a two-sided test. Hypotheses must be set up *before* observing the data. If they are not, the test must be two-sided.

4.3.6 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is $\alpha = 0.05$. However, it is often 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. smaller than 0.05). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0. Many would use $\alpha = 0.01$ in this situation.

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. We will discuss this particular case in greater detail in Section 4.6.

Significance levels should reflect consequences of errors
The significance level selected for a test should reflect the consequences associated with Type 1 and Type 2 Errors.

Example 4.35 A medical machine manufacturer is considering a higher quality but more expensive supplier for parts in making an MRI. They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier. Is there good reason to modify the significance level in such a hypothesis test?

The null hypothesis is that the more expensive parts last no more than 12% longer while the alternative is that they do last more than 12% longer. This decision is just one of the many regular factors that have a marginal impact on the MRI and the company financial health. A significance level of 0.05 seems reasonable since neither a Type 1 or Type 2 error should be dangerous or (relatively) much more expensive since the machine's accuracy won't be affected.

Example 4.36 Now consider that the same MRI manufacturer is considering a slightly more expensive supplier for parts related to safety not longevity. If the durability of the machine's components is shown to be better than the current supplier, they will switch manufacturers. Is there good reason to modify the significance level in such an evaluation?

The null hypothesis would be that the suppliers' parts are equally reliable and equally accurate in detection. Because safety is involved, the MRI machine company should be eager to switch to the slightly more expensive manufacturer (reject H_0) even if the evidence of increased safety and effectiveness is only moderately strong. A slightly larger significance level, such as $\alpha = 0.10$, might be appropriate.

Guided Practice 4.37 A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken and still detects the most common injuries at the same level with a fixed part. The part is replaced only if we are extremely certain it is broken based on a series of measurements. Identify appropriate hypotheses for this test (in plain language) and suggest an appropriate significance level.[49]

4.4 Examining the Central Limit Theorem Closer (Special Topic)

Looking back to 4.2.5, we discovered that the normal model for the sample mean tends to be very good when the sample consists of at least 30 independent observations and the

[49] Here the null hypothesis is that the part is not broken, and the alternative is that it is broken. If we don't have sufficient evidence to reject H_0, we would not replace the part. It sounds like failing to fix the part if it is broken (H_0 false, H_A true) is not very problematic, and replacing the part is expensive. Thus, we should require very strong evidence against H_0 before we replace the part. Choose a small significance level, such as $\alpha = 0.01$.

population data are not strongly skewed. The Central Limit Theorem provides the theory that allows us to make this assumption.

Central Limit Theorem, informal definition

The distribution of $\bar{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three theoretical cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *log-normal* distribution. Recall the properties of these distributions from Chapter **??**. These distributions are shown in the top panels of Figure 4.20. The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the center and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

⊙ **Guided Practice 4.38** Examine the distributions in each row of Figure 4.20. What do you notice about the normal approximation for each sampling distribution as the sample size becomes larger?[50]

● **Example 4.39** Would the normal approximation be good in all applications where the sample size is at least 30?

Not necessarily. For example, the normal approximation for the log-normal example is questionable for a sample size of 30. Generally, the more skewed a population distribution or the more common the frequency of outliers, the larger the sample required to guarantee the distribution of the sample mean is nearly normal.

TIP: With larger n, the sampling distribution of $\bar{x}$ becomes more normal

As the sample size increases, the normal model for $\bar{x}$ becomes more reasonable. We can also relax our condition on skew when the sample size is very large.

We discussed in Section 4.1.3 that the sample standard deviation, s, could be used as a substitute of the population standard deviation, σ, when computing the standard error. This estimate tends to be reasonable when $n \geq 30$. We will encounter alternative distributions for smaller sample sizes in Chapters **??** and **??**.

● **Example 4.40** Figure 4.21 shows a histogram of 50 observations. These represent the number of patient visits in a hospital for 50 consecutive days relative to their

[50] The normal approximation becomes better as larger samples are used.

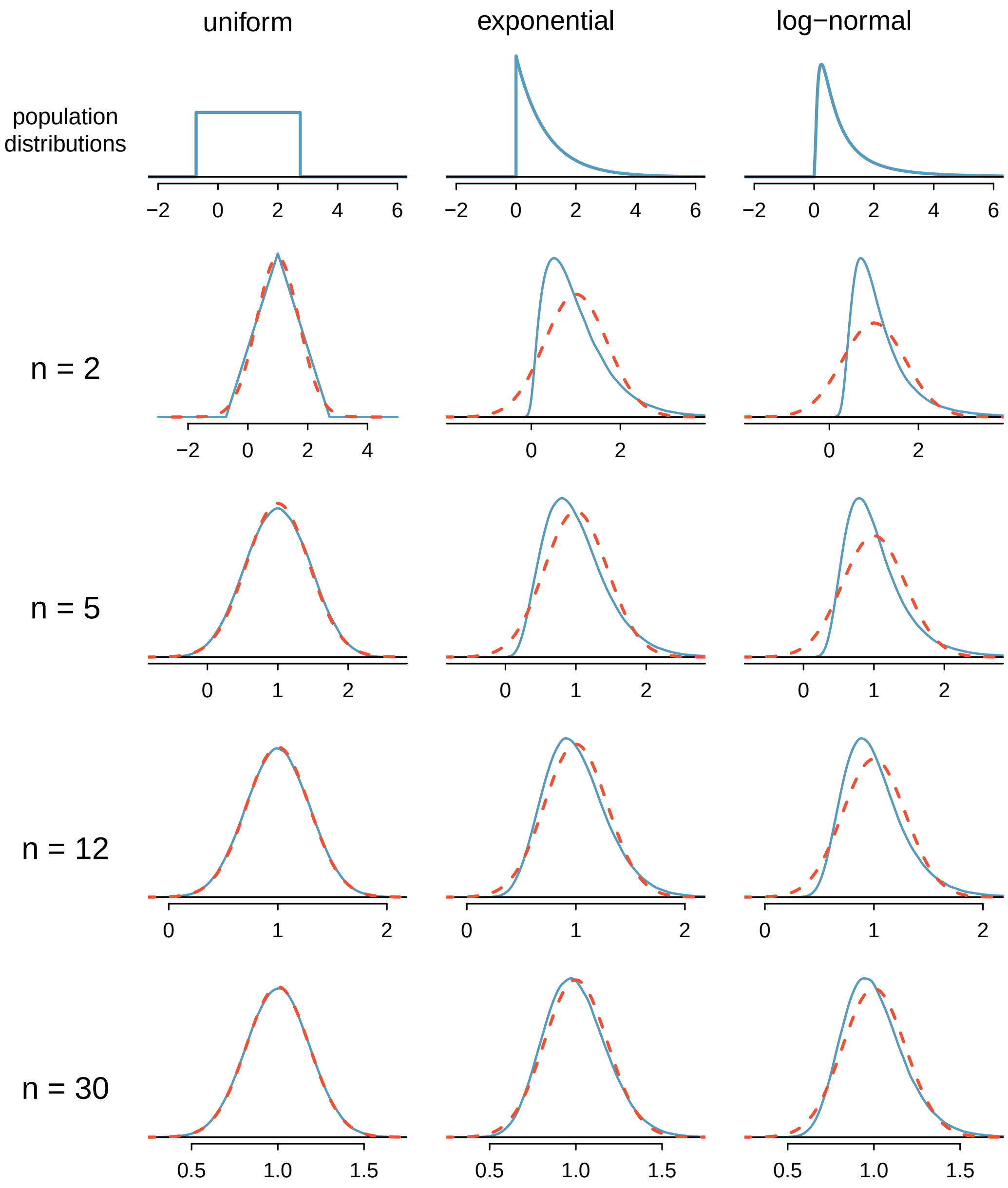


Figure 4.20: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

average rate of 5000 patient visits. Can the normal approximation be applied to the sample mean, 90.69?

We should consider each of the required conditions.

(1) These are referred to as **time series data**, because the data arrived in a particular sequence. Time series data generally deals with, you guessed it, time! If there are a lot of patients in the hospital one day, it may influence how many patients there are the day after. During the flu season, patient visits might be at an all time high since many people are sick but also the time per visit is also extremely low. To make the assumption of independence we should perform careful checks on such data. While the supporting analysis is not shown, no evidence was found to indicate the observations are not independent on a whole.

(2) The sample size is 50, satisfying the sample size condition.

(3) There are two outliers, one very extreme, which suggests the data are very strongly skewed or very distant outliers may be common for this type of data. Outliers can play an important role and affect the distribution of the sample mean and the estimate of the standard error.

Since we should be skeptical of the independence of observations and the very extreme upper outlier poses a challenge, we should not use the normal model for the sample mean of these 50 observations. If we can obtain a much larger sample, perhaps several hundred observations over a longer period of time, then the concerns about skew and outliers would no longer apply.

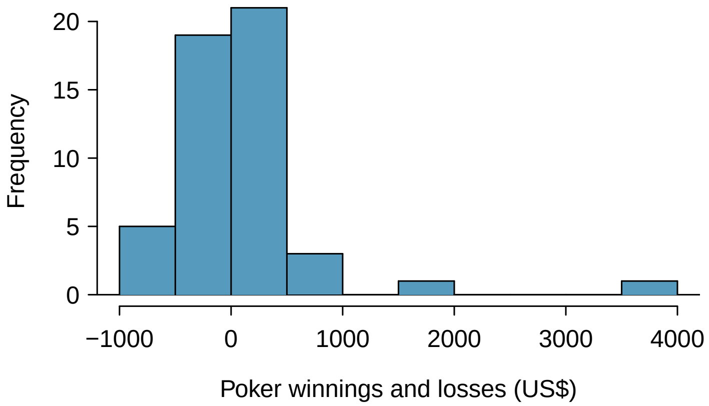


Figure 4.21: Sample distribution of total patient visits net of 5,000 visits. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

Caution: Examine data structure when considering independence

Some data sets are collected in such a way that they have a natural underlying structure between observations, e.g. when observations occur consecutively. Be especially cautious about independence assumptions regarding such data sets.

Caution: Watch out for strong skew and outliers

Strong skew is often identified by the presence of clear outliers. If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for $\bar{x}$. There are no simple guidelines for what sample size is big enough for all situations, so proceed with caution when working in the presence of strong skew or more extreme outliers.

4.5 Inference for other estimators

The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample proportions closely resembles the normal distribution when the sample size is sufficiently large. In this section, we introduce a number of examples where the normal approximation is reasonable for the point estimate. Chapters **??** and **??** will revisit each of the point estimates you see in this section along with some other new statistics.

We make another important assumption about each point estimate encountered in this section: the estimate is unbiased. A point estimate is **unbiased** if the sampling distribution of the estimate is centered at the parameter it estimates. A biased point estimate on the other hand can always be too high or estimates always too low. That is, an unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a “good” estimate. The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.

Finally, we will discuss the general case where a point estimate may follow some distribution other than the normal distribution. We also provide guidance about how to handle scenarios where the statistical techniques you are familiar with are insufficient for the problem at hand.

4.5.1 Confidence intervals for nearly normal point estimates

In Section 4.2, we used the point estimate $\bar{x}$ with a standard error $SE_{\bar{x}}$ to create a 95% confidence interval for the population mean:

$$\bar{x} \ \pm\ 1.96 \times SE_{\bar{x}} \tag{4.41}$$

We constructed this interval by noting that the sample mean is within 1.96 standard errors of the actual mean about 95% of the time. This same logic generalizes to any unbiased point estimate that is nearly normal. We may also generalize the confidence level by using a place-holder $z^{\star}$.

General confidence interval for the normal sampling distribution case

For any unbiased point estimate, the confidence interval for a nearly normal point estimate is

$$\text{point estimate} \ \pm \ z^{\star} SE \tag{4.42}$$

We see that it is of the same form as the generalized confidence interval for the sample mean where $z^{\star}$ is selected to correspond to the confidence level, and SE represents the standard error. Remember from previously that the value $z^{\star}SE$ is called the *margin of error*.

Generally the standard error for a point estimate is estimated from the data and computed using a formula. For example, the standard error for the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

In this section, we provide the computed standard error for each example and exercise without detailing where the values came from. In future chapters, you will learn to fill in these and other details for each situation.

Example 4.43 Using the `brfss.sample` data, we computed a point estimate for the average difference in weights between me and women: $\bar{x}_{\text{men}} - \bar{x}_{\text{women}} = 36.61162$ pounds. This point estimate is associated with a nearly normal distribution with SE = 0.35 pounds. What is a reasonable 95% confidence interval for the difference in gender weights?

The normal approximation is said to be valid, so we apply Equation (4.42):

$$\text{point estimate} \ \pm \ z^{\star}SE \quad \rightarrow \quad 36.61 \ \pm \ 1.96 \times 0.35 \quad \rightarrow \quad (35.91, 37.31)$$

Thus, we are 95% confident that the men were, on average, between 35.91 to 37.31 pounds heavier than women. That is, the actual average difference is plausibly between 35.91 and 37.31 pounds with 95% confidence.

Example 4.44 Does Example 4.43 guarantee that if a husband and wife both weighted themselves, the husband would weigh between 35.91 and 37.31 pounds more than the wife?

Our confidence interval says absolutely nothing about individual observations. It only makes a statement about a plausible range of values for the *average* difference between all men and women in the US.

Guided Practice 4.45 The proportion of men in the `brfss.sample` sample is $\hat{p} = 0.42$. This sample meets certain conditions that ensure $\hat{p}$ will be nearly normal, and the standard error of the estimate is $SE_{\hat{p}} = 0.05$. Create a 90% confidence interval for the proportion of participants in the BRFSS study and thus in the US who are men.[51]

[51] We use $z^{\star} = 1.65$, and apply the general confidence interval formula:

$$\hat{p} \ \pm \ z^{\star}SE_{\hat{p}} \quad \rightarrow \quad 0.42 \ \pm \ 1.65 \times 0.05 \quad \rightarrow \quad (0.3375, 0.5025)$$

Thus, we are 90% confident that between 34% and 50% are men.

4.5.2 Hypothesis testing for nearly normal point estimates

Just as the confidence interval method works with many other point estimates and we see the obvious connection between confidence intervals and hypothesis testing, it is unsurprising that we can generalize our hypothesis testing methods to new point estimates that are unbiased. Here we only consider the p-value approach, introduced in Section 4.3.2. Remember the Hypothesis testing framework from 4.3.1.

Hypothesis testing framework using the normal model

1. First write the hypotheses in plain language, then set them up in mathematical notation using the appropriate point estimate and parameter of interest.
2. State a significance level α. We generally use $\alpha = 0.05$.
3. Compute the test-statistic using the point estimate and standard error estimate.
4. Calculate the p-value by drawing a picture of the sampling distribution under H_0. Know which area you are shading to represent the correct p-value.
5. Use the p-value to evaluate your hypotheses. Write a conclusion within the context of the problem.

For point estimates other than the sampling mean which we know to be unbiased and nearly normal for $n > 30$, students need to verify conditions to ensure that the point estimate is nearly normal and unbiased so that the standard error estimate is also reasonable. This step can be done before computing the test-statistic.

⊙ **Guided Practice 4.46** A drug called sulphinpyrazone was under consideration for use in reducing the death rate in heart attack patients. To determine whether the drug was effective, a set of 1,475 patients were recruited into an experiment and randomly split into two groups: a control group that received a placebo and a treatment group that received the new drug. What would be an appropriate null hypothesis? And the alternative?[52]

We can formalize the hypotheses from Exercise 4.46 by letting $p_{control}$ and $p_{treatment}$ represent the proportion of patients who died in the control and treatment groups, respectively. Then the hypotheses can be written as

$$\begin{aligned} H_0 &: p_{control} = p_{treatment} \quad \text{(the drug doesn't work)} \\ H_A &: p_{control} > p_{treatment} \quad \text{(the drug works)} \end{aligned}$$

or equivalently,

$$\begin{aligned} H_0 &: p_{control} - p_{treatment} = 0 \quad \text{(the drug doesn't work)} \\ H_A &: p_{control} - p_{treatment} > 0 \quad \text{(the drug works)} \end{aligned}$$

Strong evidence against the null hypothesis and in favor of the alternative would correspond to an observed difference in death rates,

$$\text{point estimate} = \hat{p}_{control} - \hat{p}_{treatment}$$

[52] The skeptic's perspective is that the drug does not work at reducing deaths in heart attack patients (H_0), while the alternative is that the drug does work (H_A).

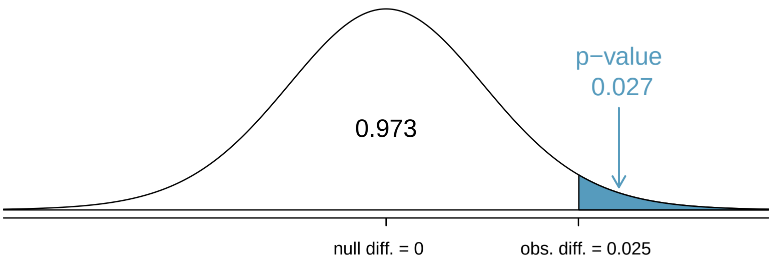


Figure 4.22: The distribution of the sample difference if the null hypothesis is true.

being larger than we would expect from chance alone. This difference in sample proportions represents a point estimate that is useful in evaluating the hypotheses.

● **Example 4.47** We want to evaluate the hypothesis setup from Exericse 4.46 using data from the actual study.[53] In the control group, 60 of 742 patients died. In the treatment group, 41 of 733 patients died. The sample difference in death rates can be summarized as

$$\text{point estimate} = \hat{p}_{control} - \hat{p}_{treatment} = \frac{60}{742} - \frac{41}{733} = 0.025$$

This point estimate is nearly normal and is an unbiased estimate of the actual difference in death rates. The standard error of this sample difference is $SE = 0.013$. Evaluate the hypothesis test at a 5% significance level: $\alpha = 0.05$.

We would like to identify the p-value to evaluate the hypotheses. If the null hypothesis is true, then the point estimate would have come from a nearly normal distribution, like the one shown in Figure 4.22. The distribution is centered at zero since $p_{control} - p_{treatment} = 0$ under the null hypothesis. Because a large positive difference provides evidence against the null hypothesis and in favor of the alternative, the upper tail has been shaded to represent the p-value. We need not shade the lower tail since this is a one-sided test: an observation in the lower tail does not support the alternative hypothesis.

The p-value can be computed by using the Z score of the point estimate and the normal probability table.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE_{\text{point estimate}}} = \frac{0.025 - 0}{0.013} = 1.92 \tag{4.48}$$

Examining Z in the normal probability table, we find that the lower unshaded tail is about 0.973. Thus, the upper shaded tail representing the p-value is

$$\text{p-value} = 1 - 0.973 = 0.027$$

Because the p-value is less than the significance level ($\alpha = 0.05$), we say the null hypothesis is implausible. That is, we reject the null hypothesis in favor of the alternative and conclude that the drug is effective at reducing deaths in heart attack patients.

[53] 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.

4.5.3 Non-normal point estimates

We may apply the ideas of confidence intervals and hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:

- the sample size is too small for the normal approximation to be valid;
- the standard error estimate may be poor; or
- the point estimate tends towards some distribution that is not the normal distribution.

For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for confidence intervals and hypothesis testing to these alternative distributions.

4.5.4 When to retreat

Statistical tools rely on conditions. When the conditions are not met, these tools are unreliable and drawing conclusions from them is treacherous. The conditions for these tools typically come in two forms.

- **The individual observations must be independent.** A random sample from less than 10% of the population ensures the observations are independent. In experiments, we generally require that subjects are randomized into groups. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.
- **Other conditions focus on sample size and skew.** For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.

Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a statistician.[54] The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Finally, we caution that there may be no inference tools helpful when considering data that include unknown biases, such as convenience samples. For this reason, there are books, courses, and researchers devoted to the techniques of sampling and experimental design. See Sections 1.3-1.5 for basic principles of data collection.

4.6 Sample size and power (special topic)

The Type 2 Error rate and the magnitude of the error for a point estimate are controlled by the sample size [55]. Real differences from the null value, even large ones, may be difficult to detect with small samples. If we take a very large sample, we might find a statistically significant difference but the magnitude might be so small that it is of no practical value. In this section we describe techniques for selecting an appropriate sample size based on these considerations.

[54] If you work at a university, then there may be campus consulting services to assist you. Alternatively, there are many private consulting firms that are also available for hire.

[55] Remember the margin of error comes from the confidence interval (point estimate ± margin of error where the margin of error $= q^{\star} \cdot SE$ for a certain confidence level)

4.6.1 Finding a sample size for a certain margin of error

Many companies are concerned about rising healthcare costs. A company may estimate certain health characteristics of its employees, such as blood pressure, to project its future cost obligations. However, it might be too expensive to measure the blood pressure of every employee at a large company, and the company may choose to take a sample instead.

● **Example 4.49** Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg. How large of a sample is necessary to estimate the average systolic blood pressure with a margin of error of 4 mmHg using a 95% confidence level?

First, we frame the problem carefully. Recall that the margin of error is the part we add and subtract from the point estimate when computing a confidence interval. Here we assume that the company has more than 30 employees and thus we can use 1.96 as the critical value for this nearly normal point estimate [56] The margin of error for a 95% confidence interval estimating a mean can be written as

$$ME_{95\%} = 1.96 \times SE = 1.96 \times \frac{\sigma_{employee}}{\sqrt{n}}$$

The challenge in this case is to find the sample size n so that this margin of error is less than or equal to 4, which we write as an inequality:

$$1.96 \times \frac{\sigma_{employee}}{\sqrt{n}} \leq 4$$

In the above equation we wish to solve for the appropriate value of n, but we need a value for $\sigma_{employee}$ before we can proceed. However, we haven't yet collected any data, so we have no direct estimate! Instead, we use the best estimate available to us: the approximate standard deviation for the U.S. population, 25. To proceed and solve for n, we substitute 25 for $\sigma_{employee}$:

$$\begin{aligned} 1.96 \times \frac{\sigma_{employee}}{\sqrt{n}} \approx 1.96 \times \frac{25}{\sqrt{n}} &\leq 4 \\ 1.96 \times \frac{25}{4} &\leq \sqrt{n} \\ \left(1.96 \times \frac{25}{4}\right)^2 &\leq n \\ 150.06 &\leq n \end{aligned}$$

This suggests we should choose a sample size of at least 151 employees. We round up because the sample size must be *greater than or equal to 150.06* to ensure a margin of error of 4.

A potentially controversial part of Example 4.49 is the use of the U.S. standard deviation for the employee standard deviation. Usually the standard deviation for the sample is not known since we haven't taken the sample just yet! In such cases, many practicing

[56] Students should verify the other assumptions as well: independence etc.

statisticians review scientific literature or market research to make an educated guess about the standard deviation to calculate the standard error.

Identify a sample size for a particular margin of error

To estimate the necessary sample size for a maximum margin of error m, we set up an equation to represent this relationship:

$$m \geq ME = q^{\star}\frac{\sigma}{\sqrt{n}}$$

where $z^{\star}$ is chosen to correspond to the desired confidence level for a nearly normal point estimate, and σ is the standard deviation associated with the population. Solve for the sample size, n.
If we believed the point estimate not to be nearly normal, use $q^{\star}$ from the T-distribution instead. However in practice, a nearly normal point estimate is used more often than not.

Sample size computations are helpful in planning data collection, and they require careful forethought. Next we consider another topic important in planning data collection and setting a sample size: the Type 2 Error rate.

4.6.2 Power and the Type 2 Error rate

Consider the following two hypotheses:

H_0: The average blood pressure of employees is the same as the national average, $\mu = 130$.

H_A: The average blood pressure of employees is different than the national average, $\mu \neq 130$.

Suppose the alternative hypothesis is actually true. Then we might like to know, what is the chance we make a Type 2 Error? That is, what is the chance we will fail to reject the null hypothesis even though we should reject it? The answer is not obvious! If the average blood pressure of the employees is 132 (just 2 mmHg from the null value), it might be very difficult to detect the difference unless we use a large sample size. On the other hand, it would be easier to detect a difference if the real average of employees was 140.

● **Example 4.50** Suppose the actual employee average is 132 and we take a sample of 100 individuals. Then the true sampling distribution of $\bar{x}$ is approximately $N(132, 2.5)$ (since $SE = \frac{25}{\sqrt{100}} = 2.5$). What is the probability of successfully rejecting the null hypothesis?

This problem can be divided into two normal probability questions. First, we identify what values of $\bar{x}$ would represent sufficiently strong evidence to reject H_0. Second, we use the hypothetical sampling distribution with center $\mu = 132$ to find the probability of observing sample means in the areas we found in the first step.

Step 1. The null distribution could be represented by $N(130, 2.5)$, the same standard deviation as the true distribution but with the null value as its center. Then we can find the two tail areas by identifying the T-statistic corresponding to the 2.5% tails

(± 1.96), and solving for x in the T-statistic equation:

$$-1.96 = T_1 = \frac{x_1 - 130}{2.5} \qquad +1.96 = T_2 = \frac{x_2 - 130}{2.5}$$
$$x_1 = 125.1 \qquad x_2 = 134.9$$

(An equally valid approach is to recognize that x_1 is $1.96 \times SE$ below the mean and x_2 is $1.96 \times SE$ above the mean to compute the values.) Figure 4.23 shows the null distribution on the left with these two dotted cutoffs.

Step 2. Next, we compute the probability of rejecting H_0 if $\bar{x}$ actually came from $N(132, 2.5)$. This is the same as finding the two shaded tails for the second distribution in Figure 4.23. We again use the T-statistic method:

$$T_{left} = \frac{125.1 - 132}{2.5} = -2.76 \qquad T_{right} = \frac{134.9 - 132}{2.5} = 1.16$$
$$area_{left} = 0.003 \qquad area_{right} = 0.123$$

The probability of rejecting the null mean, if the true mean is 132, is the sum of these areas: $0.003 + 0.123 = 0.126$.

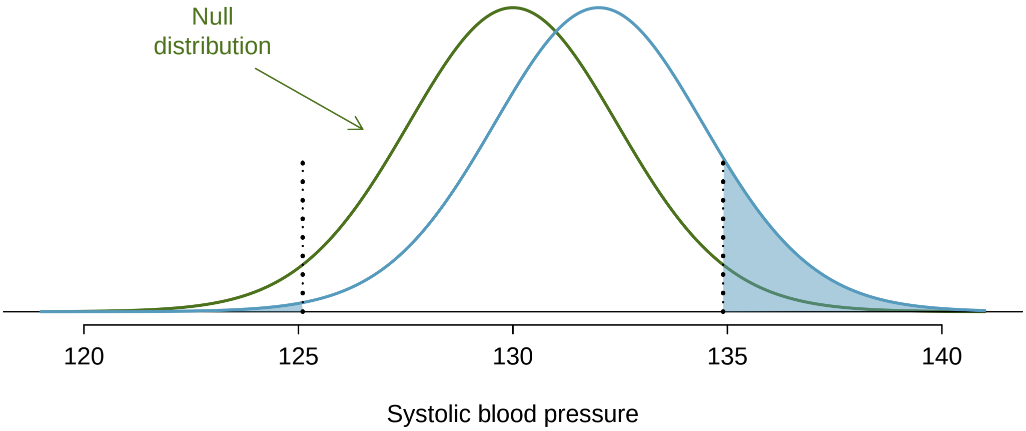


Figure 4.23: The sampling distribution of $\bar{x}$ under two scenarios. Left: $N(130, 2.5)$. Right: $N(132, 2.5)$, and the shaded areas in this distribution represent the power of the test.

The probability of rejecting the null hypothesis is called the **power**. The power varies depending on what we suppose the truth might be. In Example 4.50, the difference between the null value and the supposed true mean was relatively small, so the power was also small: only 0.126. However, when the truth is far from the null value, where we use the standard error as a measure of what is far, the power tends to increase.

Guided Practice 4.51 Suppose the true sampling distribution of $\bar{x}$ is centered at 140. That is, $\bar{x}$ comes from $N(140, 2.5)$. What would the power be under this scenario? It may be helpful to draw $N(140, 2.5)$ and shade the area representing power on Figure 4.23; use the same cutoff values identified in Example 4.50.[57]

[57] Draw the distribution $N(140, 2.5)$, then find the area below 125.1 (about zero area) and above 134.9 (about 0.979). If the true mean is 140, the power is about 0.979.

⊙ **Guided Practice 4.52** If the power of a test is 0.979 for a particular mean, what is the Type 2 Error rate for this mean?[58]

⊙ **Guided Practice 4.53** Provide an intuitive explanation for why we are more likely to reject H_0 when the true mean is further from the null value.[59]

4.6.3 Statistical significance versus practical significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected. While we still say that difference is **statistically significant**, it might not be **practically significant**.

Statistically significant differences are sometimes so minor that they are not practically relevant. This is especially important to research: if we conduct a study, we want to focus on finding a meaningful result. We don't want to spend lots of money finding results that hold no practical and applicable value.

The role of a statistician in conducting a study often includes planning the size of the study and determining the value of α. Statisticians might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. They also would obtain some reasonable estimate for the standard deviation. With these important pieces of information, a sufficiently large sample size would be chosen so that the power for the meaningful difference is perhaps 80% or 90%. While larger sample sizes may still be used, statisticians in practice might advise against using them in some cases, especially in sensitive areas of research. While we note the statistical rigor in our hypothesis testing, we must also note that many of these tests must also stand up to practical significance in the real world.

[58] The Type 2 Error rate represents the probability of failing to reject the null hypothesis. Since the power is the probability we do reject, the Type 2 Error rate will be $1 - 0.979 = 0.021$.

[59] Answers may vary a little. When the truth is far from the null value, the point estimate also tends to be far from the null value, making it easier to detect the difference and reject H_0.

4.7 Exercises

4.7.1 Variability in estimates

4.1 Identify the parameter, Part I. For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(a) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

(b) In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

(c) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

(d) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

(e) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

4.2 Identify the parameter, Part II. For each of the following situations, state whether the parameter of interest is a mean or a proportion.

(a) A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.

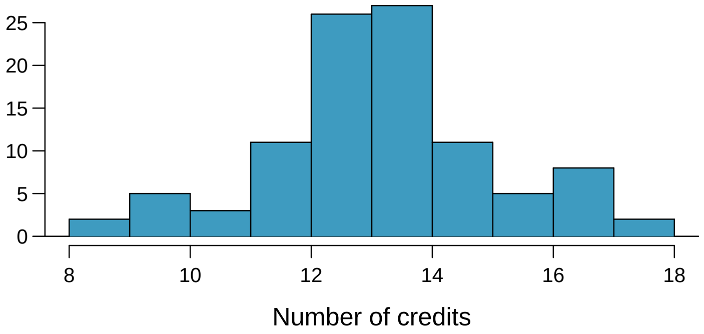
(b) A survey reports that local TV news has shown a 17% increase in revenue between 2009 and 2011 while newspaper revenues decreased by 6.4% during this time period.

(c) In a survey, high school and college students are asked whether or not they use geolocation services on their smart phones.

(d) In a survey, internet users are asked whether or not they purchased any Groupon coupons.

(e) In a survey, internet users are asked how many Groupon coupons they purchased over the last year.

4.3 College credits. A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.



Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

(a) What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

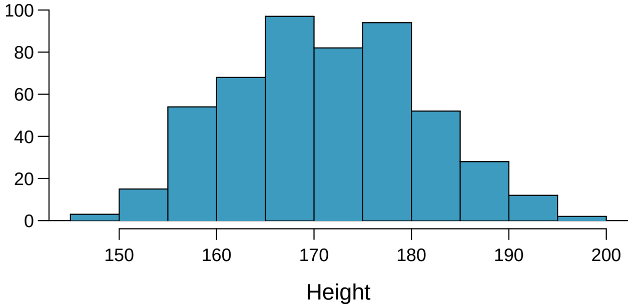
(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

(c) Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning. *Hint:* Observations farther than two standard deviations from the mean are usually considered to be unusual.

(d) The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

(e) The sample means given above are point estimates for the mean number of credits taken by all students at that college. What measures do we use to quantify the variability of this estimate? Compute this quantity using the data from the original sample.

4.4 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.[60]



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

(a) What is the point estimate for the average height of active individuals? What about the median?

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

4.5 Wireless routers. John is shopping for wireless routers and is overwhelmed by the number of available options. In order to get a feel for the average price, he takes a random sample of 75 routers and finds that the average price for this sample is $75 and the standard deviation is $25.

(a) Based on this information, how much variability should he expect to see in the mean prices of repeated samples, each containing 75 randomly selected wireless routers?

(b) A consumer website claims that the average price of routers is $80. Is a true average of $80 consistent with John's sample?

4.6 Chocolate chip cookies. Students are asked to count the number of chocolate chips in 22 cookies for a class activity. They found that the cookies on average had 14.77 chocolate chips with a standard deviation of 4.37 chocolate chips.

(a) Based on this information, about how much variability should they expect to see in the mean number of chocolate chips in random samples of 22 chocolate chip cookies?

[60] **Heinz:2003**.

(b) The packaging for these cookies claims that there are at least 20 chocolate chips per cookie. One student thinks this number is unreasonably high since the average they found is much lower. Another student claims the difference might be due to chance. What do you think?

4.7.2 Confidence intervals

4.7 Relaxing after work. The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. The survey is conducted face-to-face with an in-person interview of a randomly-selected sample of adults. One of the questions on the survey is "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" A 95% confidence interval from the 2010 GSS survey is 3.53 to 3.83 hours.[61]

(a) Interpret this interval in the context of the data.

(b) What does a 95% confidence level mean in this context?

(c) Suppose the researchers think a 90% confidence level would be more appropriate for this interval. Will this new interval be smaller or larger than the 95% confidence interval? Assume the standard deviation has remained constant since 2010.

4.8 Mental health. Another question on the General Social Survey introduced in Exercise 4.7 is "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

(a) Interpret this interval in context of the data.

(b) What does a 95% confidence level mean in this context?

(c) Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or larger than the 95% confidence interval?

(d) If a new survey asking the same questions was to be done with 500 Americans, would the standard error of the estimate be larger, smaller, or about the same. Assume the standard deviation has remained constant since 2010.

4.9 Width of a confidence interval. Earlier in Chapter 4, we calculated the 99% confidence interval for the average age of runners in the 2012 Cherry Blossom Run as (32.7, 37.4) based on a sample of 100 runners. How could we decrease the width of this interval without losing confidence?

4.10 Confidence levels. If a higher confidence level means that we are more confident about the number we are reporting, why don't we always report a confidence interval with the highest possible confidence level?

4.11 Waiting at an ER, Part I. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning for those statements you identify as false.

(a) This confidence interval is not valid since we do not know if the population distribution of the ER wait times is nearly normal.

(b) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

[61] **data:gss:2010**.

(c) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.

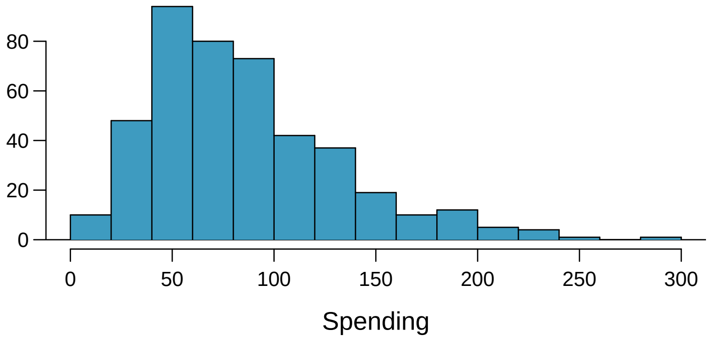
(d) 95% of such random samples would have a sample mean between 128 and 147 minutes.

(e) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

(f) The margin of error is 9.5 and the sample mean is 137.5.

(g) In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size.

4.12 Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.



(a) We are 95% confident that the average spending of these 436 American adults is between $80.31 and $89.11.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

(c) 95% of such random samples would have a sample mean between $80.31 and $89.11.

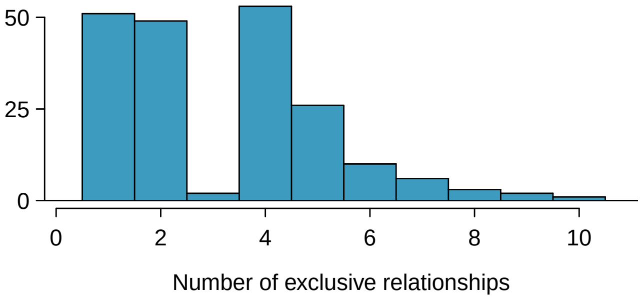
(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.

(e) A 90% confidence interval would be narrower than the 95% confidence interval.

(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

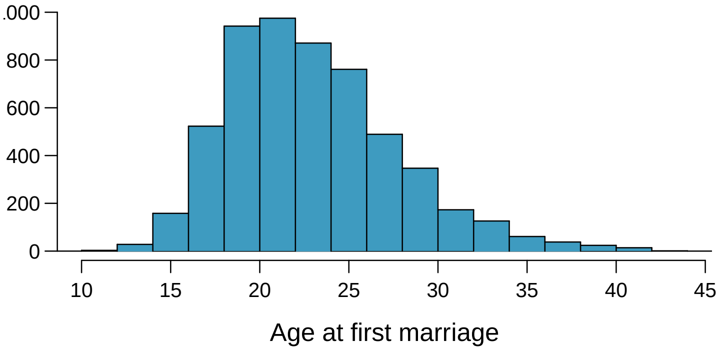
(g) The margin of error for the reported interval is 4.4.

4.13 Exclusive relationships. A survey was conducted on 203 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about the number of exclusive relationships they have been in. The histogram below shows the distribution of the data from this sample. The sample average is 3.2 with a standard deviation of 1.97.



Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.

4.14 Age at first marriage, Part I. The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram below shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72.[62]



Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.

4.7.3 Hypothesis testing

4.15 Identify hypotheses, Part I. Write the null and alternative hypotheses in words and then symbols for each of the following situations.

(a) 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. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night?

(b) Employers at a firm are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non-business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity decreases during March Madness.

[62] **data:nsfg:2010**.

4.16 Identify hypotheses, Part II. Write the null and alternative hypotheses in words and using symbols for each of the following situations.

(a) Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?

(b) Based on the performance of those who took the GRE exam between July 1, 2004 and June 30, 2007, the average Verbal Reasoning score was calculated to be 462. In 2011 the average verbal score was slightly higher. Do these data provide convincing evidence that the average GRE Verbal Reasoning score has changed since 2004?[63]

4.17 Online communication. A study suggests that the average college student spends 2 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and find that on average they spent 3.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} < 2 \textit{ hours}$$
$$H_A : \bar{x} > 3.5 \textit{ hours}$$

4.18 Age at first marriage, Part II. Exercise 4.14 presents the results of a 2006 - 2010 survey showing that the average age of women at first marriage is 23.44. Suppose a researcher believes that this value has increased in 2012, but he would also be interested if he found a decrease. Below is how he set up his hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} = 23.44 \textit{ years old}$$
$$H_A : \bar{x} > 23.44 \textit{ years old}$$

4.19 Waiting at an ER, Part II. Exercise 4.11 provides a 95% confidence interval for the mean waiting time at an emergency room (ER) of (128 minutes, 147 minutes).

(a) A local newspaper claims that the average waiting time at this ER exceeds 3 hours. What do you think of this claim?

(b) The Dean of Medicine at this hospital claims the average wait time is 2.2 hours. What do you think of this claim?

(c) Without actually calculating the interval, determine if the claim of the Dean from part (b) would be considered reasonable based on a 99% confidence interval?

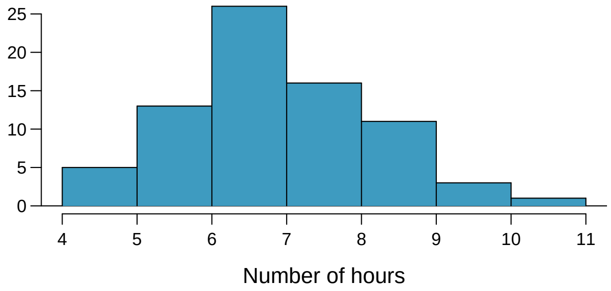
4.20 Thanksgiving spending, Part II. Exercise 4.12 provides a 95% confidence interval for the average spending by American adults during the six-day period after Thanksgiving 2009: ($80.31, $89.11).

(a) A local news anchor claims that the average spending during this period in 2009 was $100. What do you think of this claim?

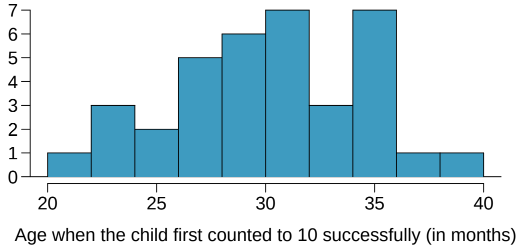
(b) Would the news anchor's claim be considered reasonable based on a 90% confidence interval? Why or why not?

4.21 Ball bearings. A manufacturer claims that bearings produced by their machine last 7 hours on average under harsh conditions. A factory worker randomly samples 75 ball bearings, and records their lifespans under harsh conditions. He calculates a sample mean of 6.85 hours, and the standard deviation of the data is 1.25 working hours. The following histogram shows the distribution of the lifespans of the ball bearings in this sample. Conduct a formal hypothesis test of this claim. Make sure to check that relevant conditions are satisfied.

[63] **webpage:GRE**.



4.22 Gifted children, Part I. Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the distribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.[64]



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

(b) Suppose you read on a parenting website that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children first count to 10 successfully is different than the general average of 32 months. Use a significance level of 0.10.

(c) Interpret the p-value in context of the hypothesis test and the data.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

4.23 Waiting at an ER, Part III. The hospital administrator mentioned in Exercise 4.11 randomly selected 64 patients and measured the time (in minutes) between when they checked in to the ER and the time they were first seen by a doctor. The average time is 137.5 minutes and the standard deviation is 39 minutes. He is getting grief from his supervisor on the basis that the wait times in the ER increased greatly from last year's average of 127 minutes. However, the administrator claims that the increase is probably just due to chance.

(a) Are conditions for inference met? Note any assumptions you must make to proceed.

(b) Using a significance level of $\alpha = 0.05$, is the change in wait times statistically significant? Use a two-sided test since it seems the supervisor had to inspect the data before he suggested an increase occurred.

(c) Would the conclusion of the hypothesis test change if the significance level was changed to $\alpha = 0.01$?

[64] **Graybill:1994**.

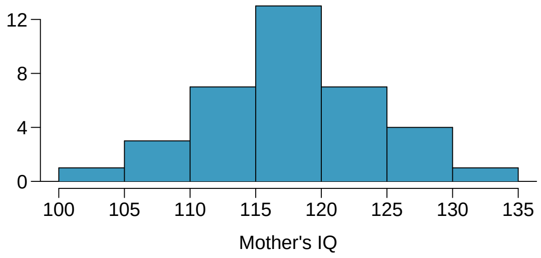
4.24 Gifted children, Part II. Exercise 4.22 describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.

n	36
min	101
mean	118.2
sd	6.5
max	131

(a) Perform a hypothesis test to evaluate if these data provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.



4.25 Nutrition labels. The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a 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? We have verified the independence, sample size, and skew conditions are satisfied.

4.26 Find the sample mean. You are given the following hypotheses: H_0: $\mu = 34$, H_A: $\mu > 34$. We know that the sample standard deviation is 10 and the sample size is 65. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

4.27 Testing for Fibromyalgia. A patient named Diana was diagnosed with Fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn't initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better.

(a) Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.

(b) What is a Type 1 error in this context?

(c) What is a Type 2 error in this context?

(d) How would these errors affect the patient?

4.28 Testing for food safety. A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.

(a) Write the hypotheses in words.

(b) What is a Type 1 error in this context?

(c) What is a Type 2 error in this context?

(d) Which error is more problematic for the restaurant owner? Why?

(e) Which error is more problematic for the diners? Why?

(f) As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.

4.29 Errors in drug testing. Suppose regulators monitored 403 drugs last year, each for a particular adverse response. For each drug they conducted a single hypothesis test with a significance level of 5% to determine if the adverse effect was higher in those taking the drug than those who did not take the drug; the regulators ultimately rejected the null hypothesis for 42 drugs.

(a) Describe the error the regulators might have made for a drug where the null hypothesis was rejected.

(b) Describe the error regulators might have made for a drug where the null hypothesis was not rejected.

(c) Suppose the vast majority of the 403 drugs do not have adverse effects. Then, if you picked one of the 42 suspect drugs at random, about how sure would you be that the drug really has an adverse effect?

(d) Can you also say how sure you are that a particular drug from the 361 where the null hypothesis was not rejected does not have the corresponding adverse response?

4.30 Car insurance savings, Part I. A car insurance company advertises that customers switching to their insurance save, on average, \$432 on their yearly premiums. A market researcher at a competing insurance discounter is interested in showing that this value is an overestimate so he can provide evidence to government regulators that the company is falsely advertising their prices. He randomly samples 82 customers who recently switched to this insurance and finds an average savings of \$395, with a standard deviation of \$102.

(a) Are conditions for inference satisfied?

(b) Perform a hypothesis test and state your conclusion.

(c) Do you agree with the market researcher that the amount of savings advertised is an overestimate? Explain your reasoning.

(d) Calculate a 90% confidence interval for the average amount of savings of all customers who switch their insurance.

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

4.31 Happy hour. A restaurant owner is considering extending the happy hour at his restaurant since he would like to see if it increases revenue. If it does, he will permanently extend happy hour. He estimates that the current average revenue per customer is \$18 during happy hour. He runs the extended happy hour for a week and finds an average revenue of \$19.25 with a standard deviation \$3.02 based on a simple random sample of 70 customers.

(a) Are conditions for inference satisfied?

(b) Perform a hypothesis test. Suppose the customers and their buying habits this week were no different than in any other week for this particular bar. (This may not always be a reasonable assumption.)

(c) Calculate a 90% confidence interval for the average revenue per customer.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

(e) If your hypothesis test and confidence interval suggest a significant increase in revenue per customer, why might you still not recommend that the restaurant owner extend the happy hour based on this criterion? What may be a better measure to consider?

4.32 Speed reading, Part I. A company offering online speed reading courses claims that students who take their courses show a 5 times (500%) increase in the number of words they can read in a minute without losing comprehension. A random sample of 100 students yielded an average increase of 415% with a standard deviation of 220%. Is there evidence that the company's claim is false?

(a) Are conditions for inference satisfied?

(b) Perform a hypothesis test evaluating if the company's claim is reasonable or if the true average improvement is less than 500%. Make sure to interpret your response in context of the hypothesis test and the data. Use $\alpha = 0.025$.

(c) Calculate a 95% confidence interval for the average increase in the number of words students can read in a minute without losing comprehension.

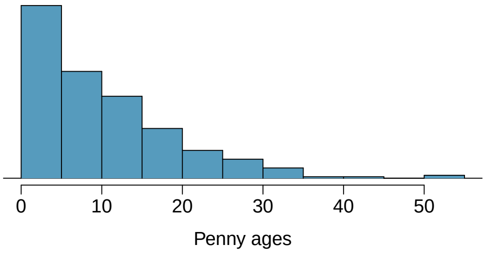
(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

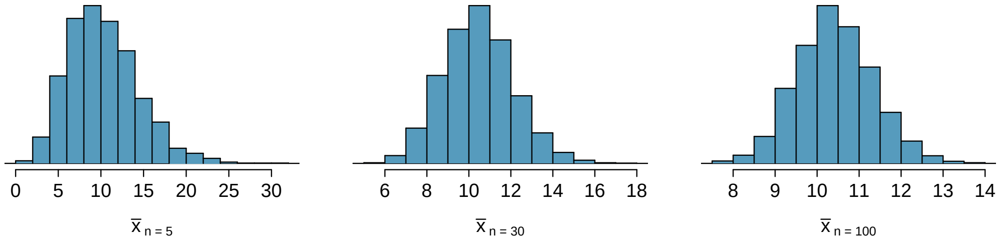
4.7.4 Examining the Central Limit Theorem

4.33 Ages of pennies, Part I. The histogram below shows the distribution of ages of pennies at a bank.

(a) Describe the distribution.

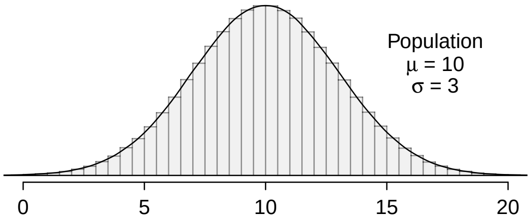
(b) Sampling distributions for means from simple random samples of 5, 30, and 100 pennies is shown in the histograms below. Describe the shapes of these distributions and comment on whether they look like what you would expect to see based on the Central Limit Theorem.

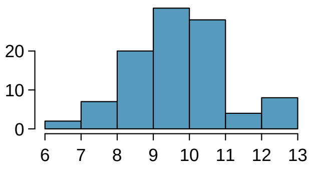




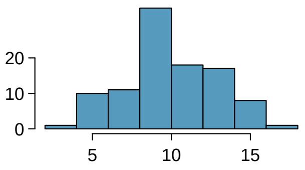
4.34 Ages of pennies, Part II. The mean age of the pennies from Exercise 4.33 is 10.44 years with a standard deviation of 9.2 years. Using the Central Limit Theorem, calculate the means and standard deviations of the distribution of the mean from random samples of size 5, 30, and 100. Comment on whether the sampling distributions shown in Exercise 4.33 agree with the values you compute.

4.35 Identify distributions, Part I. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 10 and the standard deviation is 3. Also shown below is a distribution of (1) a single random sample of 100 values from this population, (2) a distribution of 100 sample means from random samples with size 5, and (3) a distribution of 100 sample means from random samples with size 25. Determine which plot (A, B, or C) is which and explain your reasoning.

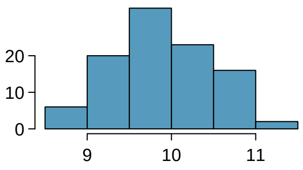




Plot A

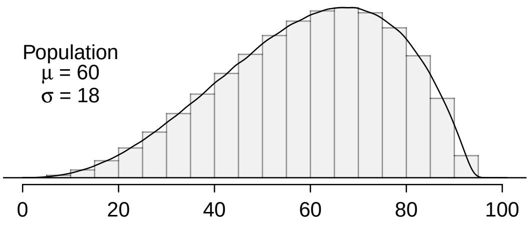


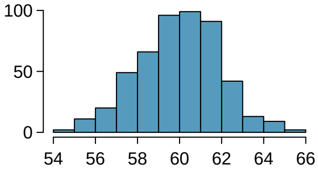
Plot B



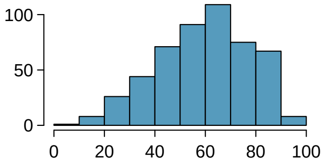
Plot C

4.36 Identify distributions, Part II. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 60 and the standard deviation is 18. Also shown below is a distribution of (1) a single random sample of 500 values from this population, (2) a distribution of 500 sample means from random samples of each size 18, and (3) a distribution of 500 sample means from random samples of each size 81. Determine which plot (A, B, or C) is which and explain your reasoning.

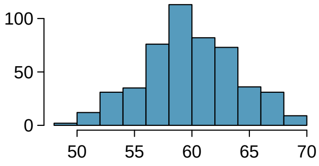




Plot A



Plot B



Plot C

4.37 Housing prices, Part I. A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

(a) Is the distribution of housing prices in Topanga symmetric, right skewed, or left skewed? *Hint:* Sketch the distribution.

(b) Would you expect most houses in Topanga to cost more or less than \$1.3 million?

(c) Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?

(d) What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

(e) How would doubling the sample size affect the standard error of the mean?

4.38 Stats final scores. Each year about 1500 students take the introductory statistics course at a large university. This year scores on the final exam are distributed with a median of 74 points, a mean of 70 points, and a standard deviation of 10 points. There are no students who scored above 100 (the maximum score attainable on the final) but a few students scored below 20 points.

(a) Is the distribution of scores on this final exam symmetric, right skewed, or left skewed?

(b) Would you expect most students to have scored above or below 70 points?

(c) Can we calculate the probability that a randomly chosen student scored above 75 using the normal distribution?

(d) What is the probability that the average score for a random sample of 40 students is above 75?

(e) How would cutting the sample size in half affect the standard error of the mean?

4.39 Weights of pennies. The distribution of weights of US pennies is approximately normal with a mean of 2.5 grams and a standard deviation of 0.03 grams.

(a) What is the probability that a randomly chosen penny weighs less than 2.4 grams?

(b) Describe the sampling distribution of the mean weight of 10 randomly chosen pennies.

(c) What is the probability that the mean weight of 10 pennies is less than 2.4 grams?

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from (a) and (c) if the weights of pennies had a skewed distribution?

4.40 CFLs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

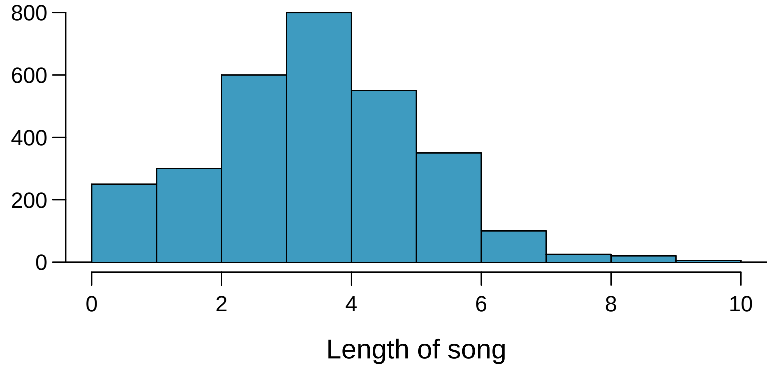
(b) Describe the distribution of the mean lifespan of 15 light bulbs.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

4.41 Songs on an iPod. Suppose an iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes.



(a) Calculate the probability that a randomly selected song lasts more than 5 minutes.

(b) You are about to go for an hour run and you make a random playlist of 15 songs. What is the probability that your playlist lasts for the entire duration of your run? *Hint:* If you want the playlist to last 60 minutes, what should be the minimum average length of a song?

(c) You are about to take a trip to visit your parents and the drive is 6 hours. You make a random playlist of 100 songs. What is the probability that your playlist lasts the entire drive?

4.42 Spray paint. Suppose the area that can be painted using a single can of spray paint is slightly variable and follows a nearly normal distribution with a mean of 25 square feet and a standard deviation of 3 square feet.

(a) What is the probability that the area covered by a can of spray paint is more than 27 square feet?

(b) Suppose you want to spray paint an area of 540 square feet using 20 cans of spray paint. On average, how many square feet must each can be able to cover to spray paint all 540 square feet?

(c) What is the probability that you can cover a 540 square feet area using 20 cans of spray paint?

(d) If the area covered by a can of spray paint had a slightly skewed distribution, could you still calculate the probabilities in parts (a) and (c) using the normal distribution?

4.7.5 Inference for other estimators

4.43 Spam mail, Part I. The 2004 National Technology Readiness Survey sponsored by the Smith School of Business at the University of Maryland surveyed 418 randomly sampled Americans, asking them how many spam emails they receive per day. The survey was repeated on a new random sample of 499 Americans in 2009.[65]

(a) What are the hypotheses for evaluating if the average spam emails per day has changed from 2004 to 2009.

(b) In 2004 the mean was 18.5 spam emails per day, and in 2009 this value was 14.9 emails per day. What is the point estimate for the difference between the two population means?

(c) A report on the survey states that the observed difference between the sample means is not statistically significant. Explain what this means in context of the hypothesis test and the data.

(d) Would you expect a confidence interval for the difference between the two population means to contain 0? Explain your reasoning.

4.44 Nearsightedness. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted.

(a) Construct hypotheses appropriate for the following question: do these data provide evidence that the 8% value is inaccurate?

(b) What proportion of children in this sample are nearsighted?

(c) Given that the standard error of the sample proportion is 0.0195 and the point estimate follows a nearly normal distribution, calculate the test statistic (the Z statistic).

(d) What is the p-value for this hypothesis test?

(e) What is the conclusion of the hypothesis test?

4.45 Spam mail, Part II. The National Technology Readiness Survey from Exercise 4.43 also asked Americans how often they delete spam emails. 23% of the respondents in 2004 said they delete their spam mail once a month or less, and in 2009 this value was 16%.

(a) What are the hypotheses for evaluating if the proportion of those who delete their email once a month or less (or never) has changed from 2004 to 2009?

(b) What is the point estimate for the difference between the two population proportions?

(c) A report on the survey states that the observed decrease from 2004 to 2009 is statistically significant. Explain what this means in context of the hypothesis test and the data.

(d) Would you expect a confidence interval for the difference between the two population proportions to contain 0? Explain your reasoning.

4.46 Unemployment and relationship problems. A USA Today/Gallup poll conducted between 2010 and 2011 asked a group of unemployed and underemployed Americans if they have had major problems in their relationships with their spouse or another close family member as a result of not having a job (if unemployed) or not having a full-time job (if underemployed). 27%

[65] **webpage:spam**.

of the 1,145 unemployed respondents and 25% of the 675 underemployed respondents said they had major problems in relationships as a result of their employment status.

(a) What are the hypotheses for evaluating if the proportions of unemployed and underemployed people who had relationship problems were different?

(b) The p-value for this hypothesis test is approximately 0.35. Explain what this means in context of the hypothesis test and the data.

4.7.6 Sample size and power

4.47 Which is higher? In each part below, there is a value of interest and two scenarios (I and II). For each part, report if the value of interest is larger under scenario I, scenario II, or whether the value is equal under the scenarios.

(a) The standard error of $\bar{x}$ when $s = 120$ and (I) n = 25 or (II) n = 125.

(b) The margin of error of a confidence interval when the confidence level is (I) 90% or (II) 80%.

(c) The p-value for a Z statistic of 2.5 when (I) n = 500 or (II) n = 1000.

(d) The probability of making a Type 2 error when the alternative hypothesis is true and the significance level is (I) 0.05 or (II) 0.10.

4.48 True or false. Determine if the following statements are true or false, and explain your reasoning. If false, state how it could be corrected.

(a) If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

(b) Decreasing the significance level (α) will increase the probability of making a Type 1 error.

(c) Suppose the null hypothesis is $\mu = 5$ and we fail to reject H_0. Under this scenario, the true population mean is 5.

(d) If the alternative hypothesis is true, then the probability of making a Type 2 error and the power of a test add up to 1.

(e) With large sample sizes, even small differences between the null value and the true value of the parameter, a difference often called the effect size, will be identified as statistically significant.

(f) A cutoff of $\alpha = 0.05$ is the ideal value for all hypothesis tests.

4.49 Car insurance savings, Part II. The market researcher from Exercise 4.30 collected data about the savings of 82 customers at a competing car insurance company. The mean and standard deviation of this sample are \$395 and \$102, respectively. He would like to conduct another survey but have a margin of error of no more than \$10 at a 99% confidence level. How large of a sample should he collect?

4.50 Speed reading, Part II. A random sample of 100 students who took online speed reading courses from the company described in Exercise 4.32 yielded an average increase in reading speed of 415% and a standard deviation of 220%. We would like to calculate a 95% confidence interval for the average increase in reading speed with a margin of error of no more than 15%. How many students should we sample?

4.51 Waiting at the ER, Part IV. Exercise 4.23 introduced us to a hospital where ER wait times were being analyzed. The previous year's average was 128 minutes. Suppose that this year's average wait time is 135 minutes.

(a) Provide the hypotheses for this situation in plain language.

(b) If we plan to collect a sample size of $n = 64$, what values could $\bar{x}$ take so that we reject H_0? Suppose the sample standard deviation from the earlier exercise (39 minutes) is the population standard deviation. You may assume that the conditions for the nearly normal model for $\bar{x}$ are satisfied.

(c) Calculate the probability of a Type 2 error.

Appendix A

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) 143,196 eligible study subjects born in Southern California between 1989 and 1993. (b) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables. (c) "Is there an association between air pollution exposure and preterm births?"

1.5 (a) 160 children. (b) Age (numerical, continuous), sex (categorical), whether they were an only child or not (categorical). (c) Research question: "Does explicitly telling children not to cheat affect their likelihood to cheat?"

1.7 (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*.

1.9 (a) Population: all births, sample: 143,196 births between 1989 and 1993 in Southern California. (b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

1.11 (a) Population: all asthma patients aged 18-69 who rely on medication for asthma treatment. Sample: 600 such patients. (b) If the patients in this sample, who are likely not randomly sampled, can be considered to be representative of all asthma patients aged 18-69 who rely on medication for asthma treatment, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.13 (a) Observation. (b) Variable. (c) Sample statistic (mean). (d) Population parameter (mean).

1.15 (a) Explanatory: number of study hours per week. Response: GPA. (b) Somewhat weak positive relationship with data becoming more sparse as the number of study hours increases. One responded reported a GPA above 4.0, which is clearly a data error. There are a few respondents who reported unusually high study hours (60 and 70 hours/week). Variability in GPA is much higher for students who study less than those who study more, which might be due to the fact that there aren't many respondents who reported studying higher hours. (c) Observational. (d) Since observational, cannot infer causation.

1.17 (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.

1.19 (a) Positive, non-linear, somewhat strong. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies, however rise in life expectancy trails off before around 80 years old. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

1.21 (a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method! (b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable. (c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)

1.23 (a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.

1.25 (a) Non-responders may have a different response to this question, e.g. parents who returned the surveys likely don't have difficulty spending time with their children. (b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might be in a lower socio- economic status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g. these people may go running because they are generally healthier and/or do other exercises.

1.27 (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population. (b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends. (d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.

1.29 No, students were not randomly sampled (voluntary sample) and the sample only contains college students at a university in Ontario.

1.31 (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Sex: man, woman.

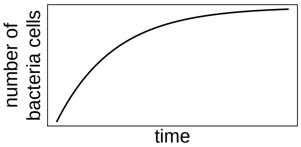
1.33 (a) Exam performance. (b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise). (c) Since the researchers want to ensure equal gender representation, sex will be a blocking variable.

1.35 Need randomization and blinding. One possible outline: (1) Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!) (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double- blind experiment. (Answers may vary.)

1.37 (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes, gender. (d) Yes, single blind since the patients were blinded to the treatment they received. (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

1.39 (a) 1: linear. 3: nonlinear.
(b) 4: linear. (c) 2.

1.41



1.43 (a) Population mean, $\mu_{2007} = 52$; sample mean, $\bar{x}_{2008} = 58$. (b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.

1.45 Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

1.47 (a) Dist 2 has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13. (b) Dist 1 has a higher mean since $-20 > -40$, and Dist 2 has a higher standard deviation since -40 is farther away from the rest of the data than -20. (c) Dist 2 has a higher mean since all values in this distribution are higher than those in Dist 1, but both distribution have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean since they're both centered at 300, but Dist 2 has a higher standard deviation since the observations are farther from the mean than in Dist 1.

1.49 (a) Q1 $\approx$ 5, median $\approx$ 15, Q3 $\approx$ 35 (b) Since the distribution is right skewed, we would expect the mean to be higher than the median.

1.51 (a) About 30. (b) Since the distribution is right skewed the mean is higher than the median. (c) Q1: between 15 and 20, Q3: between 35 and 40, IQR: about 20. (d) Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles. Upper fence: Q3 + 1.5 × IQR = $37.5 + 1.5 \times 20 = 67.5$; Lower fence: Q1 - 1.5 × IQR $= 17.5 + 1.5 \times 20 = -12.5$; The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

1.53 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.

1.55 (a) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR. (b) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR. (c) The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

1.57 No, we would expect this distribution to be right skewed. There are two reasons for this: (1) there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV), (2) the standard deviation of the distribution is very large compared to the mean.

1.59 The statement "50% of Facebook users have over 100 friends" means that the median number of friends is 100, which is lower than the mean number of friends (190), which suggests a right skewed distribution for the number of friends of Facebook users.

1.61 (a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers. (b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

1.63 (a) The distribution is unimodal and symmetric with a mean of about 25 minutes and a standard deviation of about 5 minutes. There does not appear to be any counties with unusually high or low mean travel times. Since the distribution is already unimodal and symmetric, a log transformation is not necessary. (b) Answers will vary. There are pockets of longer travel time around DC, Southeastern NY, Chicago, Minneapolis, Los Angeles, and many other big cities. There is also a large section of shorter average commute times that overlap with farmland in the Midwest. Many farmers' homes are adjacent to their farmland, so their commute would be brief, which may explain why the average commute time for these counties is relatively low.

1.65 (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.

1.67 The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.

1.69 (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True. (iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True.
(b) Proportion of all patients who had cardiovascular problems: $\frac{7{,}979}{227{,}571} \approx 0.035$
(c) The expected number of heart attacks in the rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7{,}979}{227{,}571} \approx 2370$.
(d) (i) H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance. H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems. (ii) A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that rosiglitazone increases the risk of such problems. (iii) In the actual study, we observed 2,593 cardiovascular events in the rosiglitazone group. In the 1,000 simulations under the independence model, we observed somewhat less than 2,593 in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

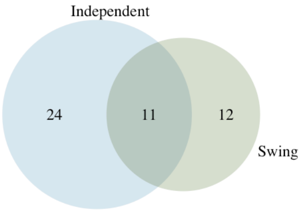
2 Probability

2.1 (a) False. These are independent trials. (b) False. There are red face cards. (c) True. A card cannot be both a face card and an ace.

2.3 (a) 10 tosses. Fewer tosses mean more variability in the sample fraction of heads, meaning there's a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

2.5 (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

2.7 (a) No, there are voters who are both independent and swing voters.
(b)



(c) Each Independent voter is either a swing voter or not. Since 35% of voters are Independents and 11% are both Independent and swing voters, the other 24% must not be swing voters. (d) 0.47. (e) 0.53. (f) P(Independent) × P(swing) = 0.35 × 0.23 = 0.08, which does not equal P(Independent and swing) = 0.11, so the events are dependent.

2.9 (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c). (b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related. (c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

2.11 (a) $0.16 + 0.09 = 0.25$. (b) $0.17 + 0.09 = 0.26$. (c) Assuming that the education level of the husband and wife are independent: $0.25 \times 0.26 = 0.065$. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

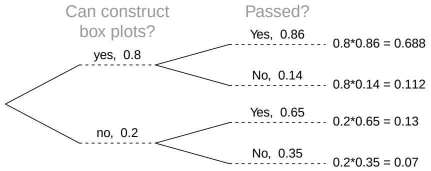
2.13 (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) Invalid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

2.15 (a) No, but we could if A and B are independent. (b-i) 0.21. (b-ii) 0.79. (b-iii) 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed under independence from part (a). (d) 0.143.

2.17 (a) No, 0.18 of respondents fall into this combination. (b) $0.60 + 0.20 - 0.18 = 0.62$. (c) $0.18/0.20 = 0.9$. (d) $0.11/0.33 \approx 0.33$. (e) No, otherwise the answers to (c) and (d) would be the same. (f) $0.06/0.34 \approx 0.18$.

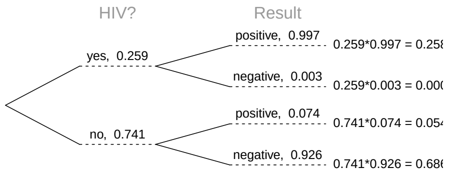
2.19 (a) No. There are 6 females who like Five Guys Burgers. (b) $162/248 = 0.65$. (c) $181/252 = 0.72$. (d) Under the assumption of a dating choices being independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (e) $(252 + 6 - 1)/500 = 0.514$.

2.21 (a)

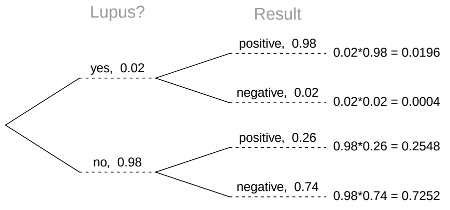


(b) 0.84

2.23 0.8247.



2.25 0.0714. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.



2.27 (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sampled from is identical in each draw.

2.29 (a) $2/9 \approx 0.22$. (b) $3/9 \approx 0.33$. (c) $\frac{3}{10} \times \frac{2}{9} \approx 0.067$. (d) No, e.g. in this exercise, removing one marble meaningfully changes the probability of what might be drawn next.

2.31 $P(^1\text{leggings}, {}^2\text{jeans}, {}^3\text{jeans}) = \frac{5}{24} \times \frac{7}{23} \times \frac{6}{23} = 0.0173$. However, the person with leggings could have come 2nd or 3rd, and these each have this same probability, so $3 \times 0.0173 = 0.519$.

2.33 (a) 13. (b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9 am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

2.35 (a) E(X) = 3.59. SD(X) = 3.37. (b) E(X) = -1.41. SD(X) = 3.37. (c) No, the expected net profit is negative, so on average you expect to lose money.

2.37 5% increase in value.

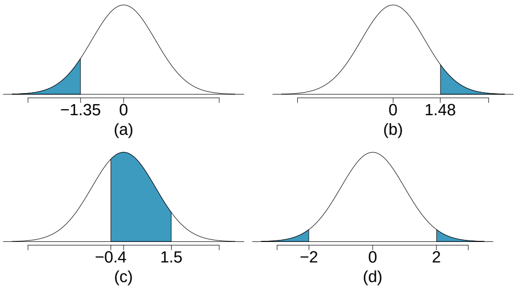
2.39 E = -0.0526. SD = 0.9986.

2.41 (a) E = \$3.90. SD = \$0.34.
(b) E = \$27.30. SD = \$0.89.

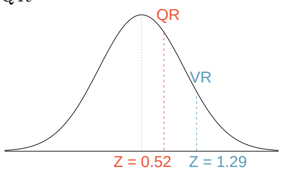
2.43 Approximate answers are OK.
(a) $(29 + 32)/144 = 0.42$. (b) $21/144 = 0.15$.
(c) $(26 + 12 + 15)/144 = 0.37$.

3 Distributions of random variables

3.1 (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.



3.3 (a) Verbal: $N(\mu = 151, \sigma = 7)$, Quant: $N(\mu = 153, \sigma = 7.67)$. (b) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$.



(c) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section. (d) She did better on the Verbal Reasoning section since her Z-score on that section was higher. (e) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (f) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ 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 (d)-(f) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

3.5 (a) $Z = 0.84$, which corresponds to approximately 160 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

3.7 (a) $Z = 1.2 \rightarrow 0.1151$.
(b) $Z = -1.28 \rightarrow 70.6°$F or colder.

3.9 (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 differ at all is because 28°C is 82.4°F, not precisely 83°F.) (d) Since $IQR = Q3 - Q1$, we first need to find $Q3$ and $Q1$ and take the difference between the two. Remember that $Q3$ is the 75^{th} and $Q1$ is the 25^{th} percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26. 86 - 23.13 = 3.73.

3.11 (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \rightarrow \sigma = \$223.88$.

3.13 $Z = 1.56 \rightarrow 0.0594$, i.e. 6%.

3.15 (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: $Z = -1.28 \rightarrow \$69.80$.

3.17 (a) 70% of the data are within 1 standard deviation of the mean, 95% are within 2 and 100% are within 3 standard deviations of the mean. Therefore, we can say that the data approximately follow the 68-95-99.7% Rule. (b) The distribution is unimodal and symmetric. The superimposed normal curve seems to approximate the distribution pretty well. The points on the normal probability plot also seem to follow a straight line. There is one possible outlier on the lower end that is apparent in both graphs, but it is not too extreme. We can say that the distribution is nearly normal.

3.19 (a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified. (b) No. There are six events under consideration. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6 and not rolling a 6, though specifying such details would be necessary.

3.21 (a) $(1 - 0.471)^2 \times 0.471 = 0.1318$. (b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$, $\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$, $\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials before a success and the standard deviation of the waiting time are higher.

3.23 (a) $0.875^2 \times 0.125 = 0.096$.
(b) $\mu = 8$, $\sigma = 7.48$.

3.25 (a) Binomial conditions are met: (1) Independent trials: In a random sample, whether or not one 18-20 year old has consumed alcohol does not depend on whether or not another one has. (2) Fixed number of trials: $n = 10$. (3) Only two outcomes at each trial: Consumed or did not consume alcohol. (4) Probability of a success is the same for each trial: $p = 0.697$. (b) 0.203. (c) 0.203. (d) 0.167. (e) 0.997.

3.27 (a) $\mu = 34.85$, $\sigma = 3.25$ (b) $Z = \frac{45-34.85}{3.25} = 3.12$. 45 is more than 3 standard deviations away from the mean, we can assume that it is an unusual observation. Therefore yes, we would be surprised. (c) Using the normal approximation, 0.0009. With 0.5 correction, 0.0015.

3.29 Want to find the probability that there will be 1,786 or more enrollees. Using the normal approximation: 0.0582. With a 0.5 correction: 0.0559.

3.31 (a) $1 - 0.75^3 = 0.5781$. (b) 0.1406. (c) 0.4219. (d) $1 - 0.25^3 = 0.9844$.

3.33 (a) Geometric distribution: 0.109. (b) Binomial: 0.219. (c) Binomial: 0.137. (d) $1 - 0.875^6 = 0.551$. (e) Geometric: 0.084. (f) Using a binomial distribution with $n = 6$ and $p = 0.75$, we see that $\mu = 4.5$, $\sigma = 1.06$, and $Z =?2.36$. Since this is not within 2 SD, it may be considered unusual.

3.35 0 wins (-\$3): 0.1458. 1 win (-\$1): 0.3936. 2 wins (+\$1): 0.3543. 3 wins (+\$3): 0.1063.

3.37 (a) $\overset{Anna}{1/5} \times \overset{Ben}{1/4} \times \overset{Carl}{1/3} \times \overset{Damian}{1/2} \times \overset{Eddy}{1/1} = 1/5! = 1/120$. (b) Since the probabilities must add to 1, there must be $5! = 120$ possible orderings. (c) $8! = 40{,}320$.

3.39 (a) 0.0804. (b) 0.0322. (c) 0.0193.

3.41 (a) Negative binomial with $n = 4$ and $p = 0.55$, where a success is defined here as a female student. The negative binomial setting is appropriate since the last trial is fixed but the order of the first 3 trials is unknown. (b) 0.1838. (c) $\binom{3}{1} = 3$. (d) In the binomial model there are no restrictions on the outcome of the last trial. In the negative binomial model the last trial is fixed. Therefore we are interested in the number of ways of orderings of the other $k - 1$ successes in the first $n - 1$ trials.

3.43 (a) Poisson with $\lambda = 75$. (b) $\mu = \lambda = 75$, $\sigma = \sqrt{\lambda} = 8.66$. (c) $Z = -1.73$. Since 60 is within 2 standard deviations of the mean, it would not generally be considered unusual. Note that we often use this rule of thumb even when the normal model does not apply. (d) Using Poisson with $\lambda = 75$: 0.0402.

4 Foundations for inference

4.1 (a) Mean. Each student reports a numerical value: a number of hours. (b) Mean. Each student reports a number, which is a percentage, and we can average over these percentages. (c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion. (d) Mean. Each student reports a number, which is a percentage like in part (b). (e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion.

4.3 (a) Mean: 13.65. Median: 14. (b) SD: 1.91. IQR: $15 - 13 = 2$. (c) $Z_{16} = 1.23$, which is not unusual since it is within 2 SD of the mean. $Z_{18} = 2.23$, which is generally considered unusual. (d) No. Point estimates that are based on samples only approximate the population parameter, and they vary from one sample to another. (e) We use the SE, which is $1.91/\sqrt{100} = 0.191$ for this sample's mean.

4.5 (a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution. (b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric. (c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = 18.2/\sqrt{45} = 2.713$. (d) The sample means will be more variable with the smaller sample size.

4.7 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:

$$45\% \pm 1.96 \times 1.2\% \quad \to \quad (42.6\%, 47.4\%)$$

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%.

4.9 (a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (b) True. Notice that the description focuses on the true population value. (c) True. If we examine the 95% confidence interval computed in Exercise **??**, 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.

4.11 (a) We are 95% confident that Americans spend an average of 1.38 to 1.92 hours per day relaxing or pursuing activities they enjoy. (b) Their confidence level must be higher as the width of the confidence interval increases as the confidence level increases. (c) The new margin of error will be smaller since as the sample size increases the standard error decreases, which will decrease the margin of error.

4.13 (a) False. Provided the data distribution is not very strongly skewed ($n = 64$ in this sample, so we can be slightly lenient with the skew), the sample mean will be nearly normal, allowing for the method normal approximation described. (b) False. Inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval. (c) True. (d) False. The confidence interval is not about a sample mean. (e) False. To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake. (f) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval. (g) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample $2^2 = 4$ times the number of people in the initial sample.

4.15 Independence: sample from $<$ 10% of population, and it is a random sample. We can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclusive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported). The sample size is at least 30. The skew is strong, but the sample is very large so this is not a concern. 90% CI: (2.97, 3.43). We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

4.17 (a) $H_0 : \mu = 8$ (On average, New Yorkers sleep 8 hours a night.)
$H_A : \mu < 8$ (On average, New Yorkers sleep less than 8 hours a night.)
(b) $H_0 : \mu = 15$ (The average amount of company time each employee spends not working is 15 minutes for March Madness.)
$H_A : \mu > 15$ (The average amount of company time each employee spends not working is greater than 15 minutes for March Madness.)

4.19 The hypotheses should be about the population mean (μ), not the sample mean. The null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. Correction:

$$H_0 : \mu = 10 \textit{ hours}$$
$$H_A : \mu > 10 \textit{ hours}$$

The one-sided test indicates that we are only interested in showing that 10 is an underestimate. Here the interest is in only one direction, so a one-sided test seems most appropriate. If we would also be interested if the data showed strong evidence that 10 was an overestimate, then the test should be two-sided.

4.21 (a) This claim does is not supported since 3 hours (180 minutes) is not in the interval. (b) 2.2 hours (132 minutes) is in the 95% confidence interval, so we do not have evidence to say she is wrong. However, it would be more appropriate to use the point estimate of the sample. (c) A 99% confidence interval will be wider than a 95% confidence interval, meaning it would enclose this smaller interval. This means 132 minutes would be in the wider interval, and we would not reject her claim based on a 99% confidence level.

4.23 $H_0 : \mu = 130$. $H_A : \mu \neq 130$. $Z = 1.39 \rightarrow$ p-value $= 0.1646$, which is larger than $\alpha = 0.05$. The data do not provide convincing evidence that the true average calorie content in bags of potato chips is different than 130 calories.

4.25 (a) Independence: The sample is random and 64 patients would almost certainly make up less than 10% of the ER residents. The sample size is at least 30. No information is provided about the skew. In practice, we would ask to see the data to check this condition, but here we will make the assumption that the skew is not very strong. (b) $H_0 : \mu = 127$. $H_A : \mu \neq 127$. $Z = 2.15 \rightarrow$ p-value $= 0.0316$. Since the p-value is less than $\alpha = 0.05$, we reject H_0. The data provide convincing evidence that the average ER wait time has increased over the last year. (c) Yes, it would change. The p-value is greater than 0.01, meaning we would fail to reject H_0 at $\alpha = 0.01$.

4.27 $Z = 1.65 = \frac{\bar{x} - 30}{10/\sqrt{70}} \rightarrow \bar{x} = 31.97$.

4.29 (a) H_0: Anti-depressants do not help symptoms of Fibromyalgia. H_A: Anti- depressants do treat symptoms of Fibromyalgia. Remark: Diana might also have taken special note if her symptoms got much worse, so a more scientific approach would have been to use a two-sided test. If you proposed a two-sided approach, your answers in (b) and (c) will be different. (b) Concluding that anti-depressants work for the treatment of Fibromyalgia symptoms when they actually do not. (c) Concluding that anti-depressants do not work for the treatment of Fibromyalgia symptoms when they actually do.

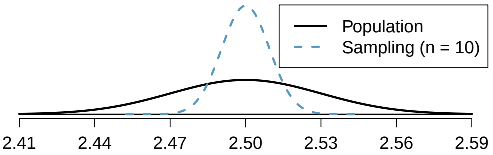
4.31 (a) Scenario I is higher. Recall that a sample mean based on less data tends to be less accurate and have larger standard errors. (b) Scenario I is higher. The higher the confidence level, the higher the corresponding margin of error. (c) They are equal. The sample size does not affect the calculation of the p- value for a given Z-score. (d) Scenario I is higher. If the null hypothesis is harder to reject (lower α), then we are more likely to make a Type 2 Error when the alternative hypothesis is true.

4.33 (a) The distribution is unimodal and strongly right skewed with a median between 5 and 10 years old. Ages range from 0 to slightly over 50 years old, and the middle 50% of the distribution is roughly between 5 and 15 years old. There are potential outliers on the higher end. (b) When the sample size is small, the sampling distribution is right skewed, just like the population distribution. As the sample size increases, the sampling distribution gets more unimodal, symmetric, and approaches normality. The variability also decreases. This is consistent with the Central Limit Theorem. (c) n = 5: $\mu_{\bar{x}} = 10.44$, $\sigma_{\bar{x}} = 4.11$; n = 30: $\mu_{\bar{x}} = 10.44$, $\sigma_{\bar{x}} = 1.68$; n = 100: $\mu_{\bar{x}} = 10.44$, $\sigma_{\bar{x}} = 0.92$. The centers of the sampling distributions shown in part (b) appear to be around 10. It is difficult to estimate the standard deviation for the sampling distribution when $n = 5$ from the histogram (since the distribution is somewhat skewed). If 1.68 is a plausible estimate for the standard deviation of the sampling distribution when $n = 30$, then using the 68-95-99.7% Rule, we would expect the values to range roughly between $10.44 \pm 3 * 1.68 = (5.4, 15.48)$, which seems to be the case. Similarly, when $n = 100$, we would expect the values to range roughly between $10.44 \pm 3 * 0.92 = (7.68, 13.2)$, which also seems to be the case.

4.35 (a) Right skewed. There is a long tail on the higher end of the distribution but a much shorter tail on the lower end. (b) Less than, as the median would be less than the mean in a right skewed distribution. (c) We should not. (d) Even though the population distribution is not normal, the conditions for inference are reasonably satisfied, with the possible exception of skew. If the skew isn't very strong (we should ask to see the data), then we can use the Central Limit Theorem to estimate this probability. For now, we'll assume the skew isn't very strong, though the description suggests it is at least moderate to strong. Use $N(1.3, SD_{\bar{x}} = 0.3/\sqrt{60})$: $Z = 2.58 \rightarrow 0.0049$. (e) It would decrease it by a factor of $1/\sqrt{2}$.

4.37 The centers are the same in each plot, and each data set is from a nearly normal distribution, though the histograms may not look very normal since each represents only 100 data points. The only way to tell which plot corresponds to which scenario is to examine the variability of each distribution. Plot B is the most variable, followed by Plot A, then Plot C. This means Plot B will correspond to the original data, Plot A to the sample means with size 5, and Plot C to the sample means with size 25.

4.39 (a) $Z = -3.33 \rightarrow 0.0004$. (b) The population SD is known and the data are nearly normal, so the sample mean will be nearly normal with distribution $N(\mu, \sigma/\sqrt{n})$, i.e. $N(2.5, 0.0095)$. (c) $Z = -10.54 \rightarrow \approx 0$. (d) See below:



(e) We could not estimate (a) without a nearly normal population distribution. We also could not estimate (c) since the sample size is not sufficient to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

4.41 (a) We cannot use the normal model for this calculation, but we can use the histogram. About 500 songs are shown to be longer than 5 minutes, so the probability is about $500/3000 = 0.167$. (b) Two different answers are reasonable. $^{Option\ 1}$Since the population distribution is only slightly skewed to the right, even a small sample size will yield a nearly normal sampling distribution. We also know that the songs are sampled randomly and the sample size is less than 10% of the population, so the length of one song in the sample is independent of another. We are looking for the probability that the total length of 15 songs is more than 60 minutes, which means that the average song should last at least $60/15 = 4$ minutes. Using $SD_{\bar{x}} = 1.63/\sqrt{15}$, $Z = 1.31 \rightarrow 0.0951$. $^{Option\ 2}$Since the population distribution is not normal, a small sample size may not be sufficient to yield a nearly normal sampling distribution. Therefore, we cannot estimate the probability using the tools we have learned so far. (c) We can now be confident that the conditions are satisfied. $Z = 0.92 \rightarrow 0.1788$.

4.43 (a) $H_0 : \mu_{2009} = \mu_{2004}$. $H_A : \mu_{2009} \neq \mu_{2004}$. (b) $\bar{x}_{2009} - \bar{x}_{2004} = -3.6$ spam emails per day. (c) The null hypothesis was not rejected, and the data do not provide convincing evidence that the true average number of spam emails per day in years 2004 and 2009 are different. The observed difference is about what we might expect from sampling variability alone. (d) Yes, since the hypothesis of no difference was not rejected in part (c).

4.45 (a) $H_0 : p_{2009} = p_{2004}$. $H_A : p_{2009} \neq p_{2004}$. (b) -7%. (c) The null hypothesis was rejected. The data provide strong evidence that the true proportion of those who once a month or less frequently delete their spam email was higher in 2004 than in 2009. The difference is so large that it cannot easily be explained as being due to chance. (d) No, since the null difference, 0, was rejected in part (c).

4.47 True. If the sample size is large, then the standard error will be small, meaning even relatively small differences between the null value and point estimate can be statistically significant.

5 Inference for numerical data

5.1 (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

5.3 (a) between 0.025 and 0.05 (b) less than 0.005 (c) greater than 0.2 (d) between 0.01 and 0.025

5.5 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 $s = 3$.

5.7 (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu < 8$ (New Yorkers sleep less 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) $0.025 <$ p-value < 0.05. 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 0.025 and 0.05. (d) Since p-value $<$ 0.05, reject H_0. The data provide strong evidence that New Yorkers sleep less than 8 hours per night on average. (e) No, as we rejected H_0.

5.9 $t^\star_{19}$ 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.

5.11 (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu < 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{20} = 0.4919$. The test statistic is $T = (4.6 - 5)/SE = -0.81$. $df = 20 - 1 = 19$. The one-tail p-value is about 0.21, which is bigger than $\alpha = 0.05$, so we do not reject H_0. That is, we do not have sufficiently strong evidence to reject Georgianna's claim.
(b) Using $SE = 0.4919$ and $t^{\star}_{df=19} = 2.093$, the confidence interval is (3.57, 5.63). We are 95% confident that the average number of years a child takes piano lessons in this city is 3.57 to 5.63 years.
(c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

5.13 If the sample is large, then the margin of error will be about $1.96 \times 100/\sqrt{n}$. We want this value to be less than 10, which leads to $n \geq 384.16$, meaning we need a sample size of at least 385 (round up for sample size calculations!).

5.15 (a) Two-sided, we are evaluating a difference, not in a particular direction. (b) Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point. (c) t-test, sample is small and population standard deviation is unknown.

5.17 (a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent. (b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group. (c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent. (d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

5.19 (a) For each observation in one data set, there is exactly one specially-corresponding observation in the other data set for the same geographic location. The data are paired. (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.) $H_A : \mu_{diff} > 0$ (Average daily high temperature in January 1, 1968 was lower than average daily high temperature in January, 2008 in the continental US.) If you chose a two-sided test, that would also be acceptable. If this is the case, note that your p-value will be a little bigger than what is reported here in part (d). (c) Independence: locations are random and represent less than 10% of all possible locations in the US. The sample size is at least 30. 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) $Z = 1.60 \rightarrow$ p-value $= 0.0548$. (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 temperature warming in the continental US. However it should be noted that the p-value is very close to 0.05. (f) Type 2 Error, 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.

5.21 (a) (-0.03, 2.23). (b) We are 90% confident that the average daily high on January 1, 2008 in the continental US was 0.03 degrees lower to 2.23 degrees higher than the average daily high on January 1, 1968. (c) No, since 0 is included in the interval.

5.23 (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. Sample size of at least 30. The skew of the differences is, at worst, slight. $Z = 2.72 \rightarrow$ p-value $= 0.0066$. 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.

5.25 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.

5.27 (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 would need to 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.

5.29 (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.

5.31 (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 Error, since we rejected H_0. (d) Yes, since p-value > 0.01, we would have failed to reject H_0.

5.33 $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ 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 can be attributed to the diet.

5.35 $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow 0.02 <$ 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.

5.37 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 P.D.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 P.D.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 P.D.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.

5.39 Difference we care about: 40. Single tail of 90%: $1.28 \times SE$. Rejection region bounds: $\pm 1.96 \times SE$ (if 5% significance level). Setting $3.24 \times SE = 40$, subbing in $SE = \sqrt{\frac{94^2}{n} + \frac{94^2}{n}}$, and solving for the sample size n gives 116 plots of land for each fertilizer.

5.41 Alternative.

5.43 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.

5.45 (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 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.

5.47 (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 $197 + 1 = 198$.

5.49 (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.

5.51 (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.

6 Inference for categorical data

6.1 (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.

6.3 (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.

6.5 (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 70% ± 8%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.

6.7 With a random sample from $< 10\%$ of the population, independence is satisfied. The success-failure condition is also satisfied. $ME = z^{\star}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56 \times 0.44}{600}} = 0.0397 \approx 4\%$

6.9 (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.

6.11 (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%.

6.13 (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).

6.15 (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.

6.17 (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 identify a soda correctly would be 0.0018.

6.19 (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.

6.21 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.

6.23 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.

6.25 (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).

6.27 (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.

6.29 (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).

6.31 (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.

6.33 Subscript $_C$ means control group. Subscript $_T$ means truck drivers. $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples, $<$ 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.

6.35 (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.

6.37 No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

6.39 (a) False. The chi-square distribution has one parameter called degrees of freedom. (b) True. (c) True. (d) False. As the degrees of freedom increases, the shape of the chi-square distribution becomes more symmetric.

6.41 (a) H_0: The distribution of the format of the book used by the students follows the professor's predictions. H_A: The distribution of the format of the book used by the students does not follow the professor's predictions. (b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$. $E_{print} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$. (c) Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, p-value > 0.3. (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.

6.43 Use a chi-squared goodness of fit test. H_0: Each option is equally likely. H_A: Some options are preferred over others. Total sample size: 99. Expected counts: (1/3) * 99 = 33 for each option. These are all above 5, so conditions are satisfied. $df = 3 - 1 = 2$ and $\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52 \rightarrow 0.02 <$ p-value < 0.05. Since the p-value is less than 5%, we reject H_0. The data provide convincing evidence that some options are preferred over others.

6.45 (a). Two-way table:

	Quit		
Treatment	Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total)\times(col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value.
(b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total)\times(col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

6.47 H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California. H_A: Opinions regarding the drilling for oil and natural gas off the coast of California has an association with earning a college degree.

$$E_{row\ 1,col\ 1} = 151.5 \qquad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \qquad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \qquad E_{row\ 3,col\ 2} = 110.5$$

Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence between observations is reasonable. Sample size: All expected counts are at least 5. $\chi^2 = 11.47$, $df = 2 \rightarrow 0.001 <$ 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.

6.49 (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

6.51 No. For a confidence interval, we check the success-failure condition using the data, and there are only 9 respondents who said bullying is no problem at all.

6.53 (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.

6.55 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.

7 Introduction to linear regression

7.1 (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.

7.3 (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.

7.5 (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).

7.7 (a) $r = -0.7 \to (4)$. (b) $r = 0.45 \to (3)$. (c) $r = 0.06 \to (1)$. (d) $r = 0.92 \to (2)$.

7.9 (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.

7.11 (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.

7.13 (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: $r = 0.636$.

7.15 (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.

7.17 In each part, we can write the husband ages as a linear function of the wife ages.
(a) $age_H = age_W + 3$.
(b) $age_H = age_W - 2$.
(c) $age_H = 2 \times age_W$.
Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, e.g. 5 women aged 26, 27, 28, 29, and 30, then find the husband ages for each wife in each part and create a scatterplot.

7.19 Correlation: no units. Intercept: kg. Slope: kg/cm.

7.21 Over-estimate. Since the residual is calculated as *observed* $-$ *predicted*, a negative residual means that the predicted value is higher than the observed value.

7.23 (a) There is a positive, very strong, linear association between the number of tourists and spending. (b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism. (d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

7.25 (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.

7.27 There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

7.29 (a) $\widehat{murder} = -29.901 + 2.559 \times poverty\%$ (b) Expected murder rate in metropolitan areas with no poverty is -29.901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in poverty, we expect murders per million to be lower on average by 2.559. (d) Poverty level explains 70.52% of the variability in murder rates in metropolitan areas. (e) $\sqrt{0.7052} = 0.8398$

7.31 (a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope.
(b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point.
(c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.

7.33 (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.

7.35 (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.

7.37 (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.

7.39 (a) $r = \sqrt{0.28} \approx -0.53$. We know the correlation is negative due to the negative association shown in the scatterplot. (b) The residuals appear to be fan shaped, indicating non-constant variance. Therefore a simple least squares fit is not appropriate for these data.

7.41 (a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million. (d) Yes, we rejected H_0 and the confidence interval does not include 0.

7.43 This is a one-sided test, so the p-value should be half of the p-value given in the regression table, which will be approximately 0. Therefore the data provide convincing evidence that poverty percentage is positively associated with murder rate.

8 Multiple and logistic regression

8.1 (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.

8.3 (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) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

8.5 (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).

8.7 Remove age.

8.9 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.)

8.11 She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

8.13 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.

8.15 (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.

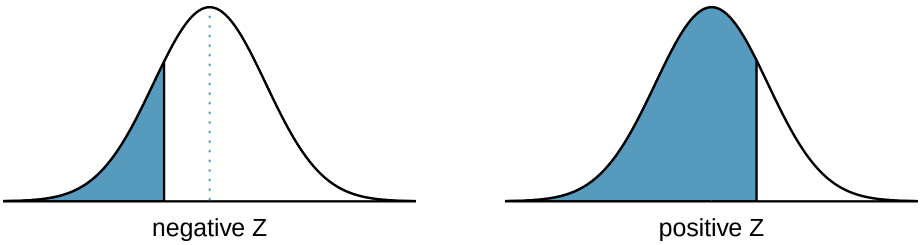
8.17 (a) The logistic model relating $\hat{p}_i$ to the predictors may be written as $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 33.5095 - 1.4207 \times sex\_male_i - 0.2787 \times skull\_width_i + 0.5687 \times total\_length_i - 1.8057 \times tail\_length_i$. Only `total_length` has a positive 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 B

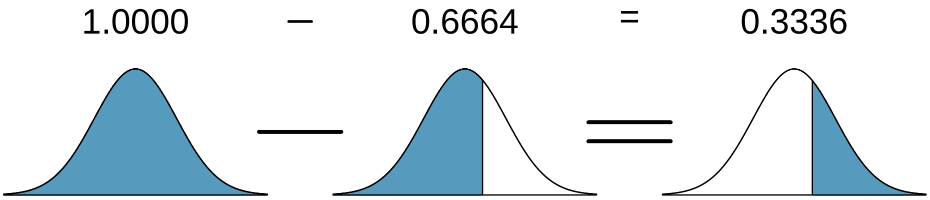
Distribution tables

B.1 Normal Probability Table

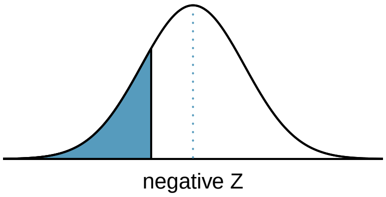
The area to the left of Z represents the percentile of the observation. The normal probability table always lists percentiles.



To find the area to the right, calculate 1 minus the area to the left.

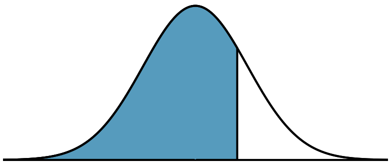


For additional details about working with the normal distribution and the normal probability table, see Section **??**, which starts on page **??**.



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	−0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	−0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	−0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	−0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.



positive Z

	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.

B.2 t-Probability Table

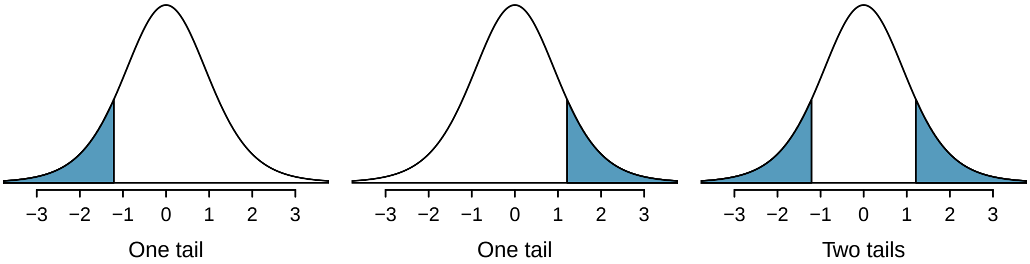


Figure B.1: Tails for the t-distribution.

	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

B.3 Chi-Square Probability Table

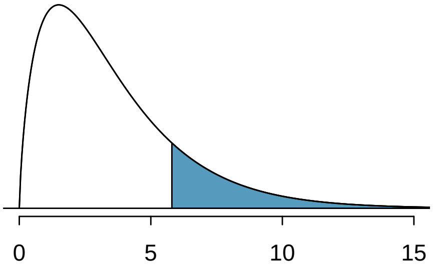


Figure B.2: Areas in the chi-square table always refer to the right tail.

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66