OpenIntro Statistics

Fourth Edition

David Diez
Data Scientist, OpenIntro

Mine Çetinkaya-Rundel
Data Scientist & Professional Educator, RStudio and Duke University

Christopher D Barr
Investment Analyst

Contents

Preface

OpenIntro Statistics is intended as a first book in statistics (or data science). This book was written for the college freshman level, but it has also been popular in high schools and graduate courses. This diverse usage highlights the book's goals: provide a rigorous introduction to applied statistics through content that is clear and concise.

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

- Statistics is an applied field with a wide range of practical applications.
- You don't have to be a math guru to learn from real, interesting data.
- Data are messy, and statistical tools are imperfect. But, when you understand the strengths and weaknesses of these tools, you can use them to learn about the world.

Textbook overview

The chapters of this book are as follows:

1. **Introduction to data.** Data structures, variables, and basic data collection techniques.
2. **Summarizing data.** Data summaries, graphics, and a teaser of inference using randomization.
3. **Probability.** Basic principles of probability.
4. **Distributions of random variables.** The normal model and other key distributions.
5. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population proportion.
6. **Inference for categorical data.** Inference for proportions and tables using the normal and chi-square distributions.
7. **Inference for numerical data.** Inference for one or two sample means using the t-distribution, statistical power for comparing two groups, and also comparisons of many means using ANOVA.
8. **Introduction to linear regression.** Regression for a numerical outcome with a single predictor. Most of this chapter could be covered after Chapter 1.
9. **Multiple and logistic regression.** Regression for numerical and categorical data using many predictors.

OpenIntro Statistics supports flexibility in choosing and ordering topics. If the main goal is to reach multiple regression (Chapter 9) as quickly as possible, then consider the following path:

- Chapter 1, Sections 2.1, and Section 2.2 for a solid introduction to data structures and statistical summaries that are used throughout the book.
- Section 4.1 for a solid understanding of the normal distribution.
- Chapter 5 to establish the core set of inference tools.
- Section 7.1 to give a foundation for the t-distribution
- Chapter 8 for establishing ideas and principles for single predictor regression.

Examples and exercises

Examples and Guided Practice throughout the textbook may be identified as follows:

EXAMPLE 0.1

(E) This is an example. When a question is asked here, where can the answer be found?

The answer can be found here, in the solution section of the example!

GUIDED PRACTICE 0.2

(G) When we think the reader should be ready to try something out or answer a question, we frame it in a Guided Practice problem. Of course, the reader may not know the answer, so we provide the full solution in a footnote.[1]

A large number of exercises are also provided at the end of each [section or chapter?] for practice or homework assignments. Solutions are given for odd-numbered exercise in Appendix **??**.

Additional resources

Video overviews, slides, statistical software labs, data sets used in the textbook, and much more are readily available at

openintro.org/os

We also have improved the ability to access data in this book through the addition of Appendix A, which provides additional information for each of the data sets used in the main text. Online guides to each of these data sets are also provided at **openintro.org/data**.

We appreciate all feedback as well as reports of any typos through the website. A short-link to report a new typo or review known typos is **openintro.org/os/typos**.

For those focused on statistics at the high school level, consider [ensure no name change] *Advanced High School Statistics*, which is a version of *OpenIntro Statistics* that has been heavily customized for high school courses and AP® Statistics by Leah Dorazio.

Acknowledgements

This project would not be possible without the passion and dedication of all those involved. The authors would like to thank the OpenIntro Staff for their involvement and ongoing contributions. We are also very grateful to the hundreds of students and instructors who have provided us with valuable feedback since we first started posting book content in 2009.

We also want to thank the many teachers who helped review this edition, including Laura Acion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, and David Rockoff. We appreciate their helpful feedback, which helped us tune the text in significant ways for the better.

[1] Guided Practice solutions are always located in a footnote.

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, and in this first chapter, we focus on both the properties of data and on the collection of data.

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

1.1 Case study: using stents to prevent strokes

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

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

Does the use of stents reduce the risk of stroke?

The researchers who asked this question conducted an experiment with 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

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

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

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

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

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

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

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

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

Figure 1.2: Descriptive statistics for the stent study.

GUIDED PRACTICE 1.1

Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all Guided Practice exercises are provided using footnotes.)[1]

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

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

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

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

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

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

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

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

1.2 Data basics

Effective organization and description of data is a first step in most analyses. This section introduces the *data matrix* for organizing data as well as some terminology about different forms of data that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

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

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

GUIDED PRACTICE 1.2

G What is the grade of the first loan in Figure 1.3? And what is the home ownership status of the borrower for that first loan? For these Guided Practice questions, you can check your answer in the footnote.[2]

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

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

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

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

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

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

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

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

GUIDED PRACTICE 1.3

G The grades for assignments, quizzes, and exams in a course are often recorded in a gradebook that takes the form of a data matrix. How might you organize grade data using a data matrix?[3]

GUIDED PRACTICE 1.4

G We consider data for 3,139 counties in the United States, which includes each county's name, the state where it resides, its population in 2017, how its population changed from 2010 to 2017, poverty rate, and six additional characteristics. How might these data be organized in a data matrix?[4]

The data described in Guided Practice 1.4 represents the `county` data set, which is shown as a data matrix in Figure 1.5. The variables are summarized in Figure 1.6.

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

[4]Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,139 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

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

Figure 1.5: Eleven rows from the `county` data set.

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

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

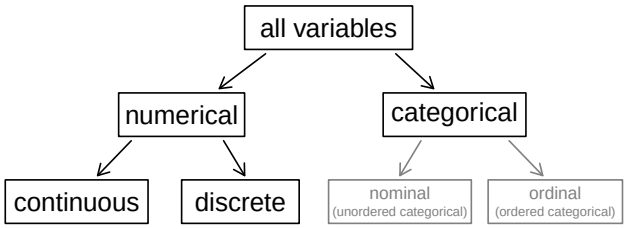


Figure 1.7: Breakdown of variables into their respective types.

1.2.2 Types of variables

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

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

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

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

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

EXAMPLE 1.5

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

(E)

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

GUIDED PRACTICE 1.6

An experiment is evaluating the effectiveness of a new drug in treating migraines. A `group` variable is used to indicate the experiment group for each patient: treatment or control. The `num_migraines` variable represents the number of migraines the patient experienced during a 3-month period. Classify each variable as either numerical or categorical?[5]

(G)

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

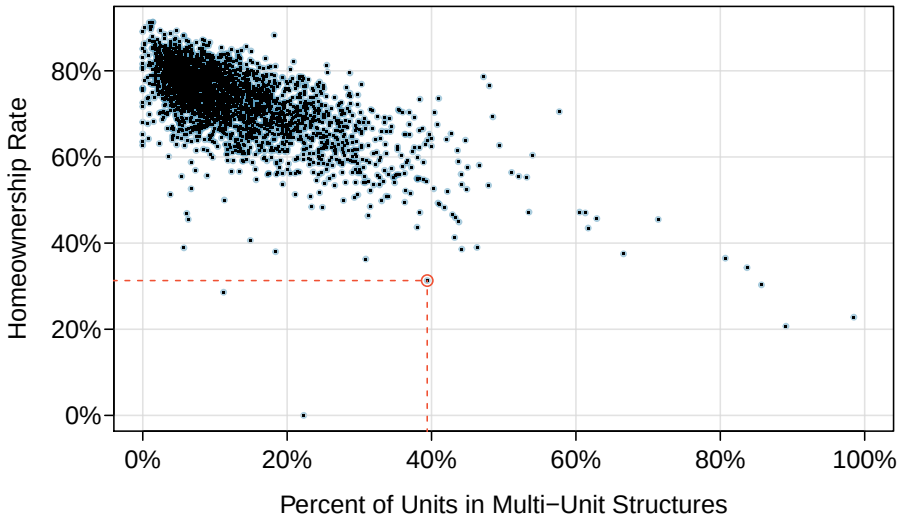


Figure 1.8: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for US counties. The highlighted dot represents Chattahoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeownership rate of 31.3%.

1.2.3 Relationships between variables

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

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

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

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

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

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

The multi-unit and homeownership rates are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables. Associated variables can also be called **dependent** variables and vice-versa.

GUIDED PRACTICE 1.7

G Examine the variables in the `loan50` data set, which are described in Table 1.4 on page 10. Create two questions about possible relationships between variables in `loan50` that are of interest to you.[6]

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

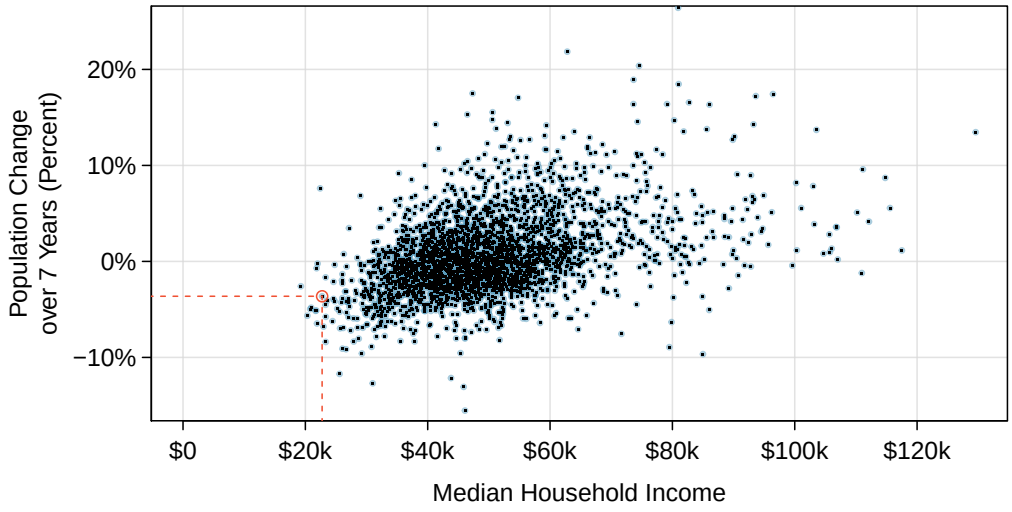


Figure 1.9: A scatterplot showing `pop_change` against `median_hh_income`. Owsley County of Kentucky, with a poverty rate of 45.2% and median household income of $22,736, is highlighted.

EXAMPLE 1.8

This example examines the relationship between population change over a 7-year period and median household income, which is visualized as a scatterplot in Figure 1.9. Are these variables associated?

(E)

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

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

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

ASSOCIATED OR INDEPENDENT, NOT BOTH

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

1.2.4 Explanatory and response variables

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

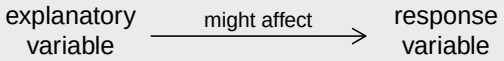
> *If a county is able to increase its median household income, does this drive an increase in its population?*

In this question, we are asking whether one variable affects another. If this is our underlying belief, then *median household income* is the **explanatory** variable and the *population change* is the **response** variable in the hypothesized relationship.[7]

[7]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,

EXPLANATORY AND RESPONSE VARIABLES

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



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

Bear in mind that the act of labeling the variables in this way does nothing to guarantee that a causal relationship exists. A formal evaluation to check whether one variable causes a change in another requires an experiment.

1.2.5 Introducing observational studies and experiments

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

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to form hypotheses about 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 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.

ASSOCIATION $\neq$ CAUSATION

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

so we avoid this language.

1.3 Sampling principles and strategies

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

1.3.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

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

GUIDED PRACTICE 1.9

G For the second and third questions above, identify the target population and what represents an individual case.[8]

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

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

ANECDOTAL EVIDENCE

Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7

[8] (2) The first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergraduate students who graduated in the last five years represent cases in the population under consideration. Each such student is an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

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

years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

1.3.3 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates. We pick samples randomly to reduce the chance we introduce biases.

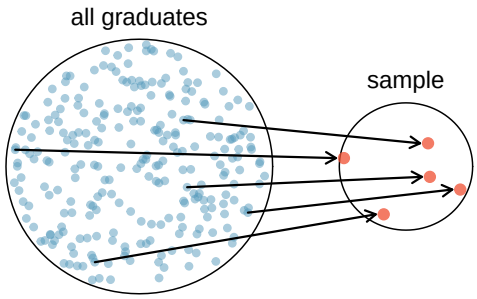


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

EXAMPLE 1.10

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?

(E)

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be a good representation of the population. When selecting samples by hand, we run the risk of picking a **biased** sample, even if their bias isn't intended.

If someone was permitted to pick and choose exactly which graduates were included in the

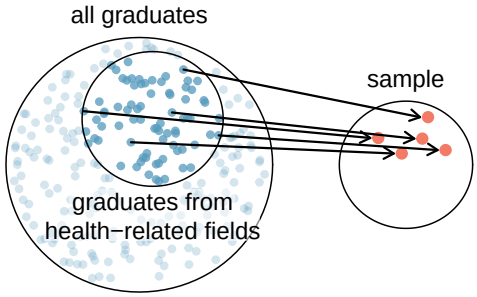


Figure 1.12: Asked to pick a sample of graduates, a nutrition major might inadvertently pick a disproportionate number of graduates from health-related majors.

sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, 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 rate** 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.

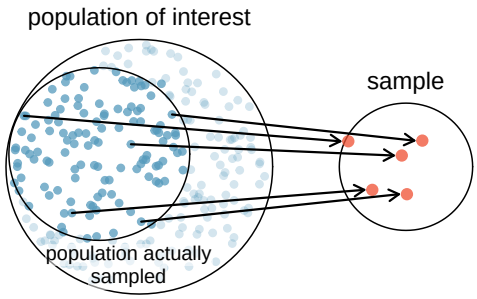


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

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 1.11

G 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?[9]

[9] 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.

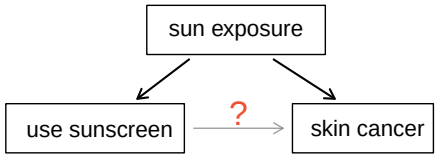
1.3.4 Observational studies

Data where no treatment has been explicitly applied (or explicitly withheld) is called **observational data**. For instance, the loan data and county data described in Section 1.2 are both examples of observational data. 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 or form hypotheses that we later check using experiments.

GUIDED PRACTICE 1.12

G 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?[10]

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**,[11] 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.

GUIDED PRACTICE 1.13

G Figure 1.8 shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest a variable that might explain the negative relationship.[12]

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 patients 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. 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 may contain both prospectively- and retrospectively-collected variables.

1.3.5 Four sampling methods

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 1.14 and 1.15 provide graphical representations of these techniques.

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

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

[12] 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.

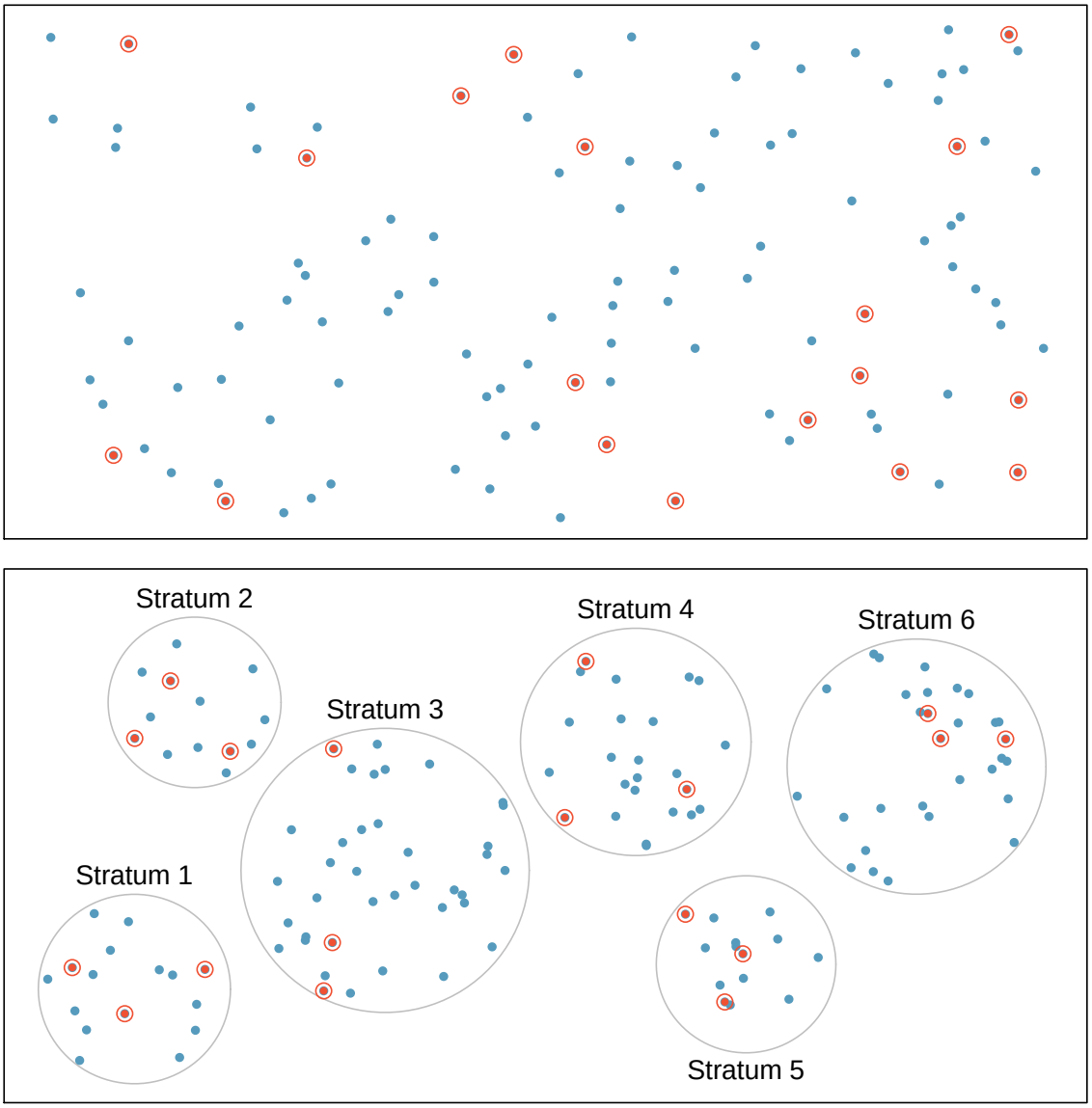


Figure 1.14: Examples of simple random and stratified sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the bottom panel, stratified sampling was used: cases were grouped into strata, then simple random sampling was employed within each stratum.

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, we could write the names of that season's several hundreds of 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.

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 1.14

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

(E) We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar, leading to more precise estimates within each group. When we combine these estimates into a single estimate for the full population, that population estimate will tend to be more precise since each individual group estimate is itself more precise.

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 techniques are typically required to analyze the data, though the methods in this book can be extended to handle such data.

EXAMPLE 1.15

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?

(E) 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 the cluster sample would still give us reliable information, even if we would need to analyze the data with slightly more advanced methods than we discuss in this book.

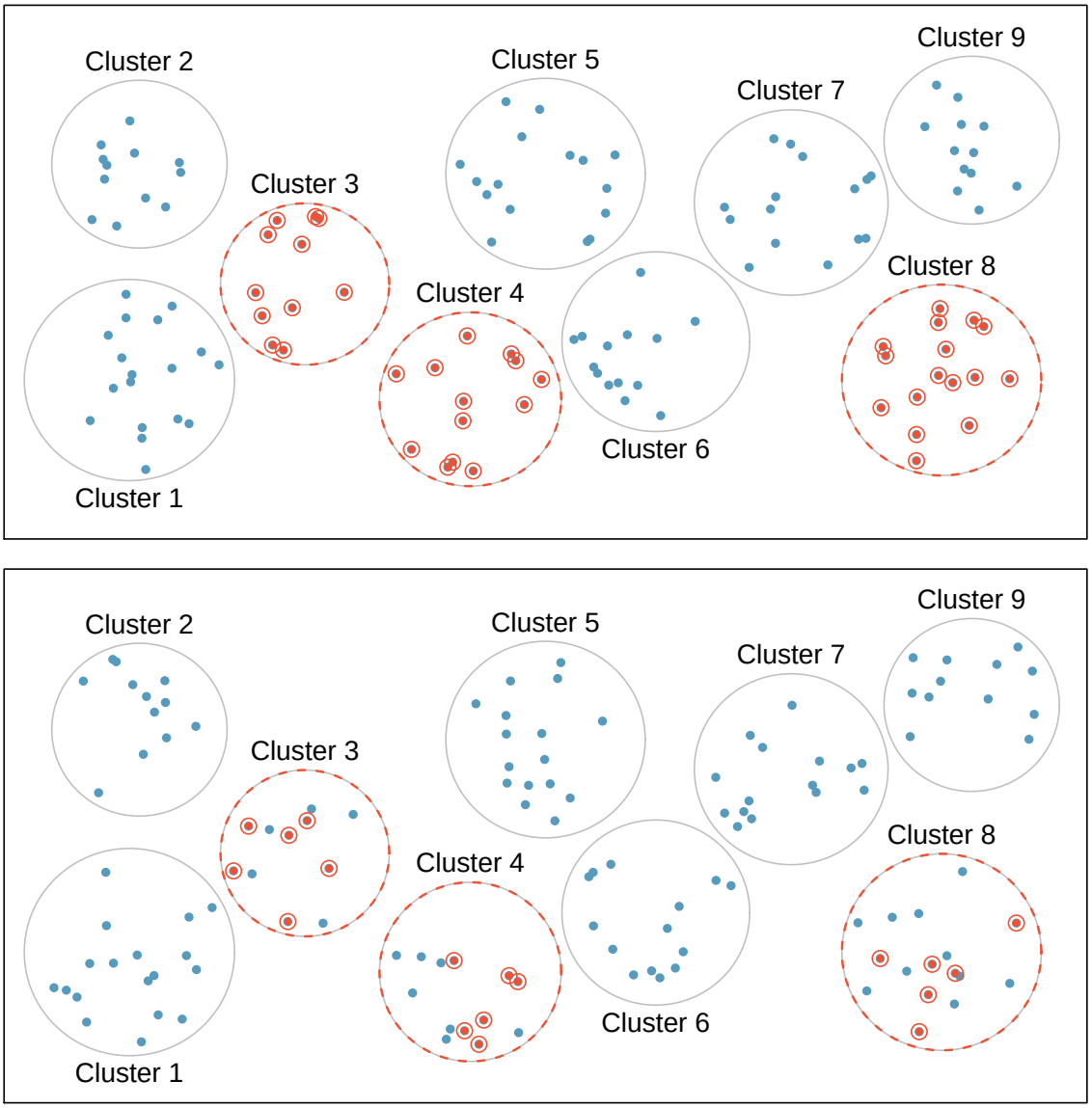


Figure 1.15: Examples of cluster and multistage sampling. In the top panel, cluster sampling was used: data were binned into nine clusters, three of these clusters were sampled, and all observations within these three cluster were included in the sample. In the bottom panel, multistage sampling was used, which differs from cluster sampling only in that we randomly select a subset of each cluster to be included in the sample rather than measuring every case in each sampled cluster.

1.4 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.

1.4.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups.[13] 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 1.16. 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.

1.4.2 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 relationship 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. 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[14] 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.

[13] This is a different concept than a *control group*, which we discuss in the second principle and in Section 1.4.2.
[14] Human subjects are often called **patients**, **volunteers**, or **study participants**.

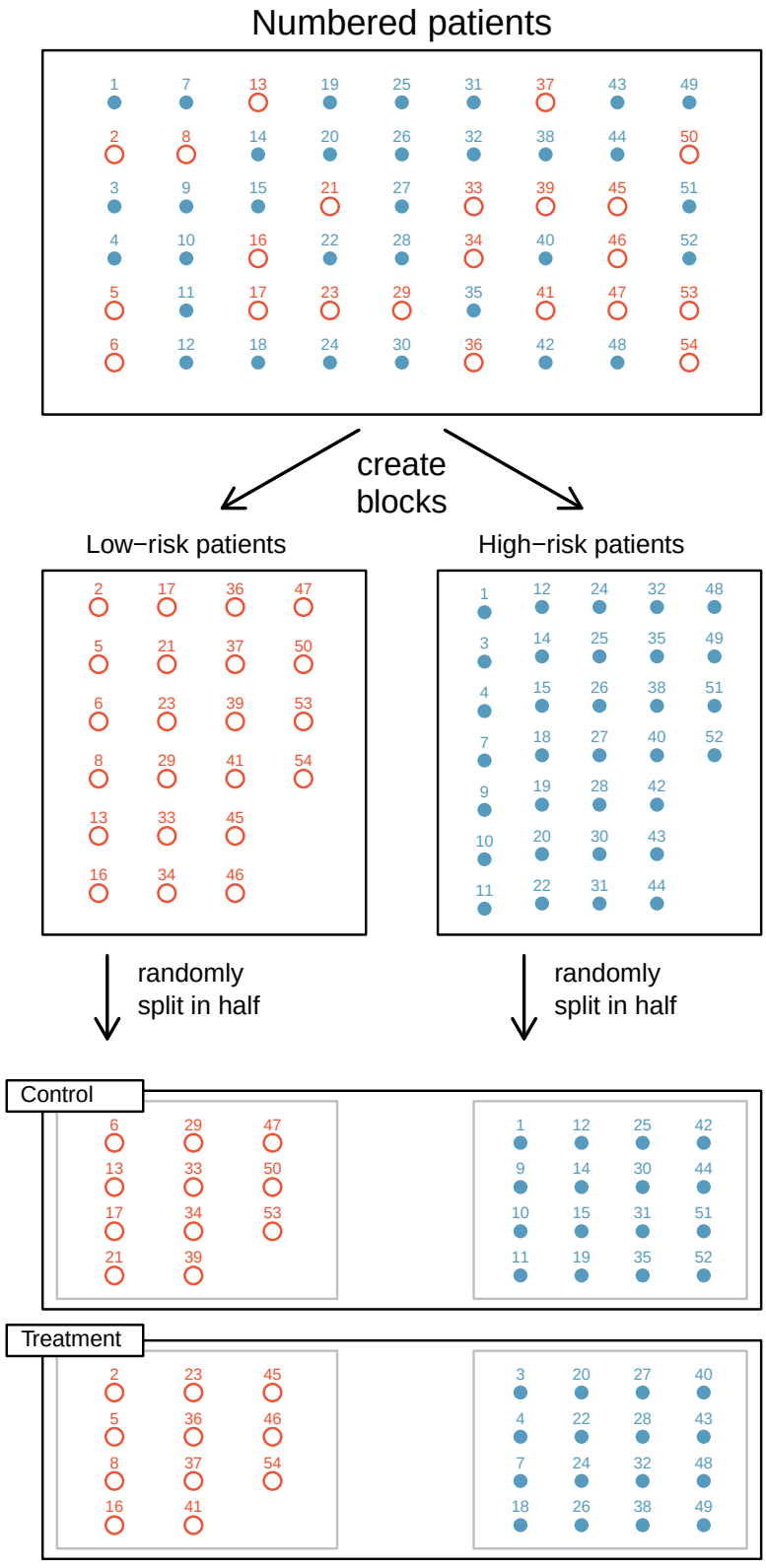


Figure 1.16: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

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

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

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[15]

GUIDED PRACTICE 1.16

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?[16]

[The next GP is new.]

GUIDED PRACTICE 1.17

For the study in Section 1.1, could the researchers have employed a placebo? If so, what would that placebo have looked like?[17]

You may have many questions about the ethics of sham surgeries to create a placebo after reading Guided Practice 1.17. These questions may have even arisen in your mind when in the general experiment context, where a possibly helpful treatment was withheld from individuals in the control group; the main difference is that a sham surgery tends to create additional risk, while withholding a treatment only maintains a person's risk.

There are always multiple viewpoints of experiments and placebos, and rarely is it obvious which is ethically "correct". For instance, is it ethical to use a sham surgery when it creates a risk to the patient? However, if we don't use sham surgeries, we may promote the use of a costly treatment that has no real effect; if this happens, money and other resources will be diverted away from other treatments that are known to be helpful. Ultimately, we cannot perfectly protect both the patients who have volunteered for the study and the patients who may benefit (or not) from the treatment in the future.

[15] 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.

[16] 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.

[17] Ultimately, can we make patients think they got treated from a surgery? In fact, we can, and some experiments use what's called a **sham surgery**. In a sham surgery, the patient does undergo surgery, but the patient does not receive the full treatment, though they will still get a placebo effect.

Chapter 2

Summarizing data

This chapter focuses on the mechanics and construction of summary statistics and graphs. In practice, we use statistical software for generating most of the summaries and graphs presented in this chapter. However since this might be your first exposure to these concepts, we take our time in this chapter to detail how to compute them. Mastery of the content presented in this chapter will be crucial for understanding the methods and techniques introduced in future chapters of the book.

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

2.1 Examining numerical data

In this section we will explore techniques for summarizing numerical variables. For example, consider the `loan_amount` variable from the `loan50` data set, which represents the loan size for all 50 loans in the data set. This variable is numerical since we can sensibly discuss the numerical difference of the size of two loans. On the other hand, area codes and zip codes are not numerical, but rather they are categorical variables.

Throughout this section and the next, we will apply these methods using the `loan50` and `county` data sets, which were introduced in Section 1.2. If you'd like to review the variables from either data set, see pages 1.3 and 1.5.

2.1.1 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 the homeownership rate against the fraction of housing units that were part of multi-unit properties (e.g. apartments) in the `county` data set. Another scatterplot is shown in Figure 2.1, comparing the total income of a borrower (`total_income`) and the amount they borrowed (`loan_amount`) for the `loan50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `loan50`, there are 50 points in Figure 2.1.

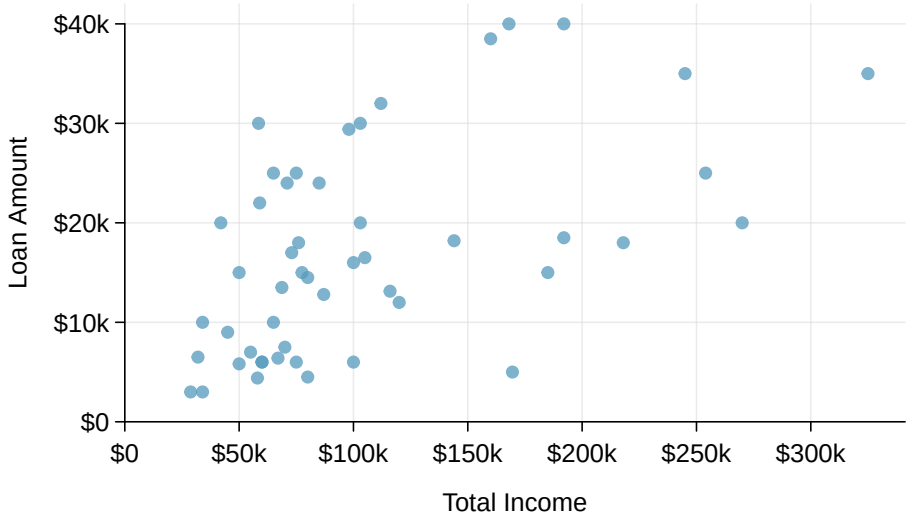


Figure 2.1: A scatterplot of `total_income` versus `loan_amount` for the `loan50` data set.

Looking at Figure 2.1, we see that there are many borrowers with an income below $100,000 on the left side of the graph, while there are a handful of borrowers with income above $250,000.

EXAMPLE 2.1

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

(E)

The relationship is evidently **nonlinear**, as highlighted by the dashed line. This is different from previous scatterplots we've seen, which show relationships that do not show much, if any, curvature in the trend.

GUIDED PRACTICE 2.2

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

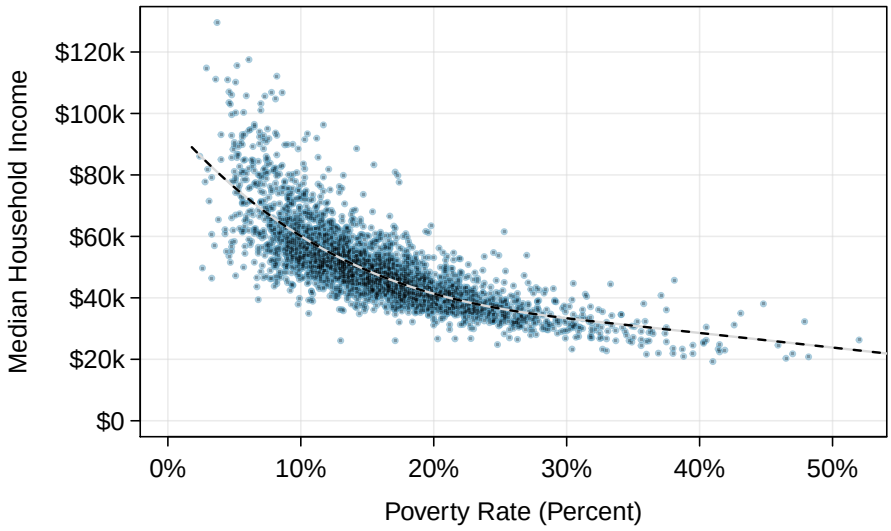


Figure 2.2: A scatterplot of the median household income against the poverty rate for the `county` data set. A statistical model has also been fit to the data and is shown as a dashed line.

G **GUIDED PRACTICE 2.3**

Describe two variables that would have a horseshoe-shaped association in a scatterplot ($\cap$ or $\frown$).[2]

2.1.2 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 interest rate of 50 loans is shown in Figure 2.3. A stacked version of this dot plot is shown in Figure 2.4.

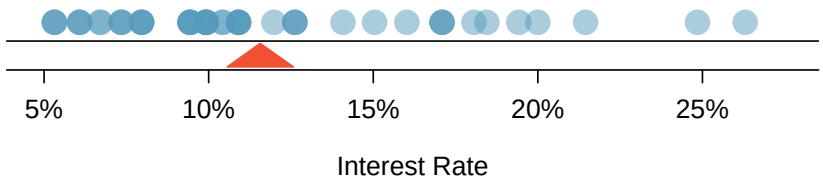


Figure 2.3: A dot plot of `interest_rate` for the `loan50` data set. The distribution's mean is shown as a red triangle.

The **mean**, often called the **average**, is a common way to measure the center of a **distribution** of data. To compute the mean interest rate, we add up all the interest rates and divide by the number of observations:

$$\bar{x} = \frac{10.90\% + 9.92\% + 26.30\% + \cdots + 6.08\%}{50} = 11.57\%$$

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `interest_rate`, and the bar over the x communicates we're looking at the

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

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

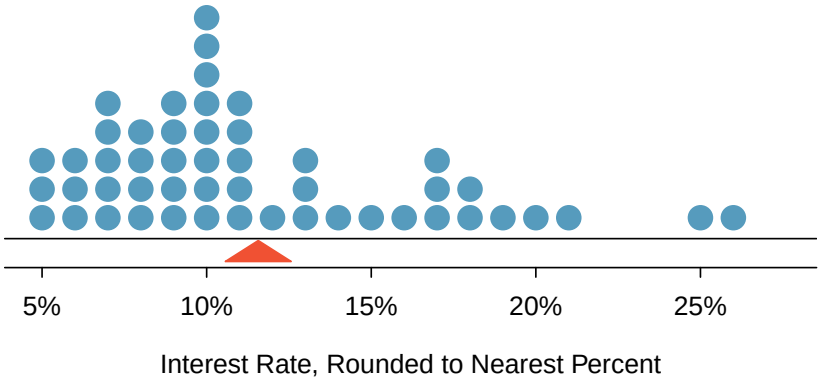


Figure 2.4: A stacked dot plot of `interest_rate` for the `loan50` data set. The rates have been rounded to the nearest percent in this plot, and the distribution's mean is shown as a red triangle.

average interest rate, which for these 50 loans was 11.57%. It is useful to think of the mean as the balancing point of the distribution, and it's shown as a triangle in Figures 2.3 and 2.4.

MEAN

The sample mean can be computed as the sum of the observed values divided by the number of observations:

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

where x_1, x_2, ..., x_n represent the n observed values.

GUIDED PRACTICE 2.4

Examine the equation for the mean. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[3]

GUIDED PRACTICE 2.5

What was n in this sample of loans?[4]

The `loan50` data set represents a sample from a larger population of loans made through Lending Club. 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. Often times it is too expensive to measure the population mean precisely, so we often estimate μ using the sample mean, $\bar{x}$.

[3] x_1 corresponds to the interest rate for the first loan in the sample (10.90%), x_2 to the second loan's interest rate (9.92%), and x_i corresponds to the interest rate for the i^{th} loan in the data set. For example, if $i = 4$, then we're examining x_4, which refers to the fourth observation in the data set.

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

EXAMPLE 2.6

The average interest rate across all loans in the population can be estimated using the sample data. Based on the sample of 50 loans, what would be a reasonable estimate of μ_x, the mean interest rate for all loans in the full data set?

(E)

The sample mean, 11.57%, provides a rough estimate of μ_x. While it's not perfect, this is our single best guess of the average interest rate of all the loans in the population under study.

In Chapter 5 and beyond, we will develop tools to characterize the accuracy of *point estimates* like the sample mean, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

The mean is useful because it allows us to rescale or standardize a metric to something more easily interpretable and comparable.

EXAMPLE 2.7

Provide a couple examples where the mean is a useful tool for making comparisons.

1. We would like to understand if a new drug is more effective at treating asthma attacks than the standard drug. A trial of 1000 adults is set up, where 500 receive the new drug, and 1000 receive a standard drug in the control group:

	New drug	Standard drug
Number of patients	500	1000
Total asthma attacks	200	300

Comparing the raw counts of 200 to 300 asthma attacks would make it appear that the new drug is better, but this is an artifact of the imbalanced group sizes. Instead, we should look at the average number of asthma attacks per patient in each group:

(E)

$$\text{New drug: } \frac{200}{500} = 0.4 \qquad \text{Standard drug: } \frac{300}{1000} = 0.3$$

The standard drug has a lower average number of asthma attacks per patient than the average in the treatment group.

2. Emilio opened a food truck last year where he sells burritos, and his business has stabilized over the last 3 months. Over that 3 month period, he has made \$11,000 while working 625 hours. Emilio's average hourly earnings provides a useful statistic for evaluating whether his venture is, at least from a financial perspective, worth it:

$$\frac{\$11000}{625 \text{ hours}} = \$17.60 \text{ per hour}$$

By knowing his average hourly wage, Emilio now has put his earnings into a standard unit that is easier to compare with many other jobs that he might consider.

EXAMPLE 2.8

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

(E)

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 \$30,861. Had we computed the *simple* mean of per capita income across counties, the result would have been just \$26,093!

Example 2.8 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 under [provide a pointer to this extra content].

2.1.3 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 `loan50` data set, we created a table of counts for the number of loans with interest rates between 5.0% and 7.5%, then the number of loans with rates between 7.5% and 10.0%, and so on. Observations that fall on the boundary of a bin (e.g. 10.00%) are allocated to the lower bin. This tabulation is shown in Figure 2.5. These binned counts are plotted as bars in Figure 2.6 into what is called a **histogram**, which resembles a more heavily binned version of the stacked dot plot shown in Figure 2.4.

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27.5%
Count	11	15	8	4	···	1

Figure 2.5: Counts for the binned `interest_rate` data.

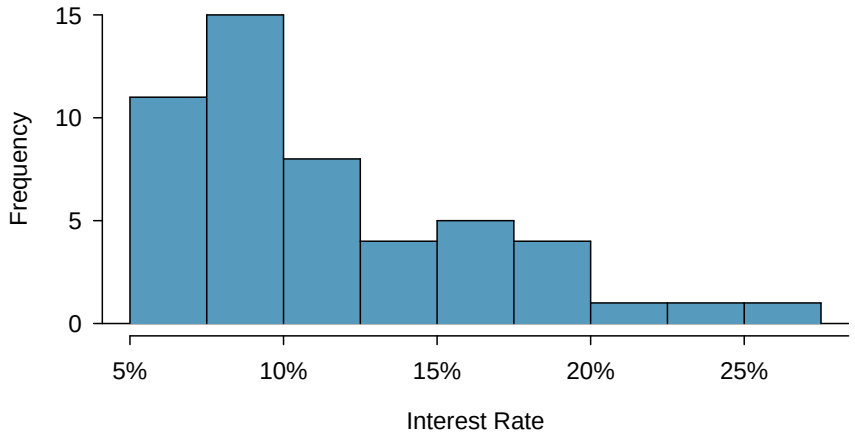


Figure 2.6: A histogram of `interest_rate`. This distribution is strongly skewed to the right.

Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common. For instance, there are many more loans with rates between 5% and 10% than loans with rates between 20% and 25% in the data set. The bars make it easy to see how the density of the data changes relative to the interest rate.

Histograms are especially convenient for understanding the shape of the data distribution. Figure 2.6 suggests that most loans have rates under 15%, while only a handful of loans have rates above 20%. When data trail off to the right in this way and has a longer right tail, the shape is said to be **right skewed**.[5]

Data sets with the reverse characteristic – a long, thinner 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.

[5]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**.

GUIDED PRACTICE 2.9

G Take a look at the dot plots in Figures 2.3 and 2.4. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[6]

GUIDED PRACTICE 2.10

G Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[7]

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. There is only one prominent peak in the histogram of `loan_amount`.

A definition of *mode* sometimes taught in math classes is the value with the most occurrences in the data set. However, for many real-world data sets, it is common to have *no* observations with the same value in a data set, making this definition impractical in data analysis.

Figure 2.7 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.

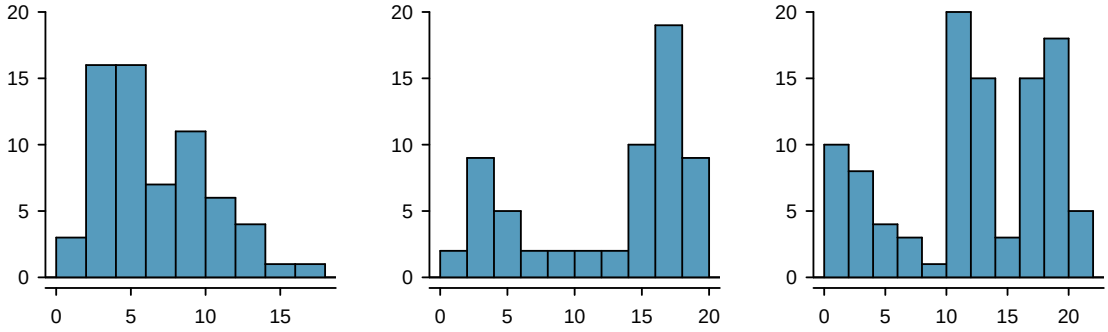


Figure 2.7: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal. Note that we've said the left plot is unimodal intentionally. This is because we are counting *prominent* peaks, not just any peak.

EXAMPLE 2.11

E Figure 2.6 reveals only one prominent mode in the interest rate. Is the distribution unimodal, bimodal, or multimodal?

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.

GUIDED PRACTICE 2.12

G Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you expect in this height data set?[8]

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 most important

[6] 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.

[7] The interest rates for individual loans.

[8] 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.

part of this examination is to better understand your data.

2.1.4 Variance and standard deviation

The mean was introduced as a method to describe the center of a data set, and 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 comprehend, 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 `interest_rate` variable:

$$\begin{aligned} x_1 - \bar{x} &= 10.90 - 11.57 = -0.67 \\ x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\ x_3 - \bar{x} &= 26.30 - 11.57 = 14.73 \\ &\vdots \\ x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49 \end{aligned}$$

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

$$\begin{aligned} s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\ &= \frac{0.45 + 2.72 + 216.97 + \cdots + 30.14}{49} \\ &= 25.52 \end{aligned}$$

We divide by $n - 1$, rather than dividing by n, when computing a sample's variance; there's some mathematical nuance here, but the end result is that doing this makes this statistic slightly more reliable and useful.

Notice that squaring the deviations does two things. First, it makes large values relatively much larger, seen by comparing $(-0.67)^2$, $(-1.65)^2$, $(14.73)^2$, and $(-5.49)^2$. Second, it gets rid of any negative signs.

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

$$s = \sqrt{25.52} = 5.05$$

While often omitted, a subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, if it is useful as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n.

VARIANCE AND STANDARD DEVIATION

The variance is 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 far the data are distributed from the mean.

The standard deviation represents the typical deviation of observations from the mean. Usually about 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 2.8 and 2.9, these percentages are not strict rules.

Like the mean, the population values for variance and standard deviation have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

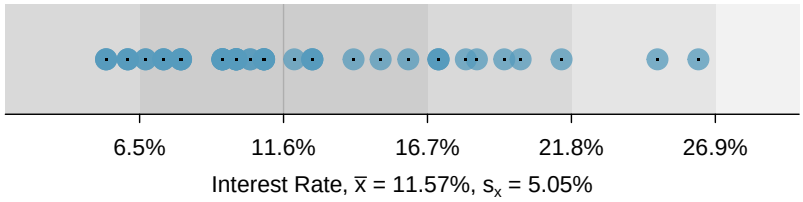


Figure 2.8: For the `interest_rate` variable, 34 of the 50 loans (68%) had interest rates within 1 standard deviation of the mean, and 48 of the 50 loans (96%) had rates within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% within 2 standard deviations, though this is far from a hard rule.

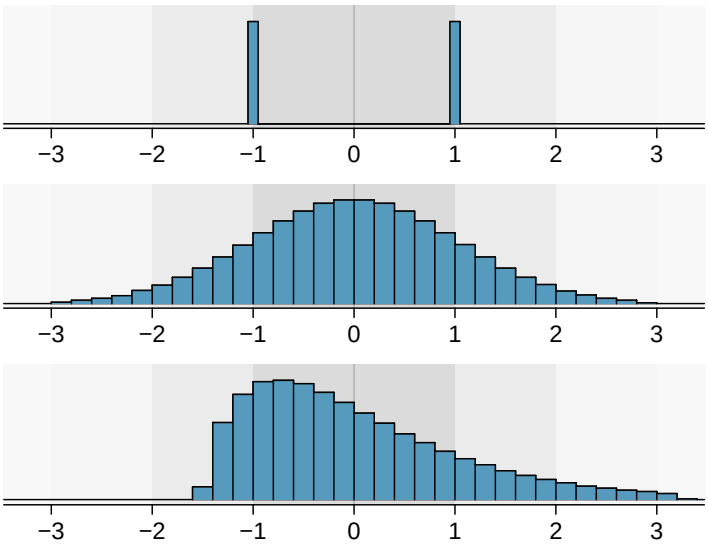


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

GUIDED PRACTICE 2.13

G On page 33, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.9 as an example, explain why such a description is important.[9]

EXAMPLE 2.14

E Describe the distribution of the `interest_rate` variable using the histogram in Figure 2.6. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context. Also note any especially unusual cases.

The distribution of interest rates is unimodal and skewed to the high end. Many of the rates fall near the mean at 11.57%, and most fall within one standard deviation (5.05%) of the mean. There are a few exceptionally large interest rates in the sample that are above 20%.

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 5 the standard deviation is used in calculations that help us understand how much a sample mean varies from one sample to the next.

2.1.5 Box plots, quartiles, and the median

A **box plot** summarizes a data set using five statistics while also plotting unusual observations. Figure 2.10 provides a vertical dot plot alongside a box plot of the `interest_rate` variable from the `loan50` data set.

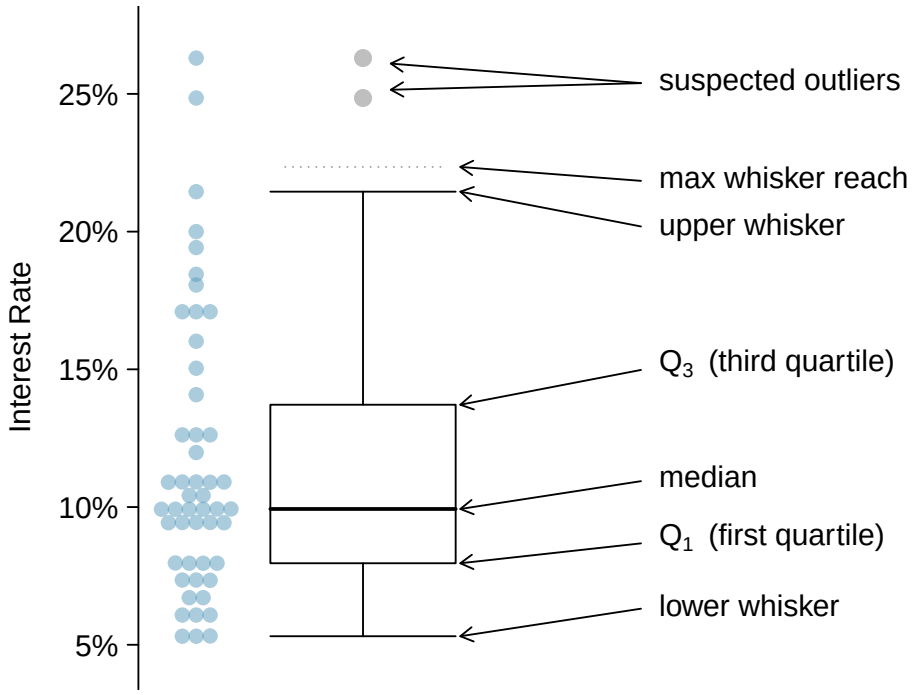


Figure 2.10: A vertical dot plot, where points have been horizontally stacked, next to a labeled box plot for the interest rates of the 50 loans.

[9] Figure 2.9 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.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 2.10 shows 50% of the data falling below the median and other 50% falling above the median. There are 50 loans 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, which happen to be the same value in this data set: $(9.93\% + 9.93\%)/2 = 9.93\%$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in such a 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 2.10, 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 tend to be. 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 2.15

(G) What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[10]

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 $1.5 \times IQR$. They capture everything within this reach. In Figure 2.10, the upper whisker does not extend to the last two 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, 5.31%, 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 lying beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of 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 interest rates of 24.85% and 26.30% 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.

Examining data for outliers serves many useful purposes, including

1. Identifying strong skew in the distribution.
2. Identifying possible data collection or data entry errors.
3. Providing insight into interesting properties of the data.

[10] 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.

GUIDED PRACTICE 2.16

Using Figure 2.10, estimate the following values for `interest_rate` in the `loan50` data set: (a) Q_1, (b) Q_3, and (c) IQR.[11]

2.1.6 Robust statistics

How are the sample statistics of the `interest_rate` data set affected by the observation, 26.3%? What would have happened if this loan had instead been only 15%? What would happen to these summary statistics if the observation at 26.3% had been even larger, say 35%? These scenarios are plotted alongside the original data in Figure 2.11, and sample statistics are computed under each scenario in Figure 2.12.

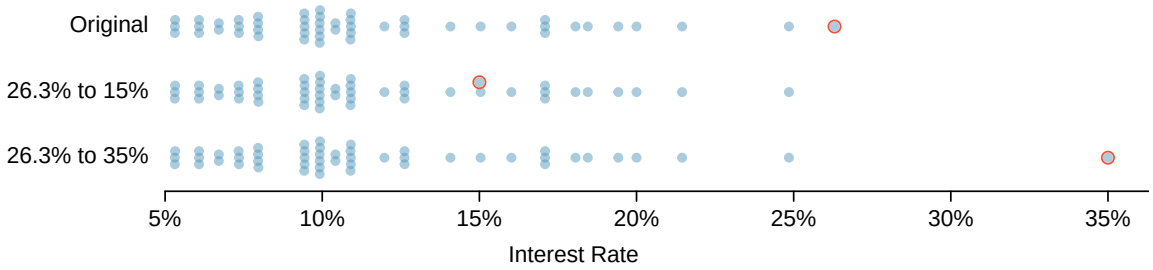


Figure 2.11: Dot plots of the original interest rate data and two modified data sets.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `interest_rate` data	9.93%	5.76%	11.57%	5.05%
move 26.3% $\rightarrow$ 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% $\rightarrow$ 35%	9.93%	5.76%	11.74%	5.68%

Figure 2.12: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change had an extreme observations from the `interest_rate` variable been different.

GUIDED PRACTICE 2.17

(a) Which is more affected by extreme observations, the mean or median? Figure 2.12 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[12]

The median and IQR are called **robust statistics** because extreme observations have little effect on their values: moving the most extreme value generally has little influence on these statistics. On the other hand, the mean and standard deviation are more heavily influenced by changes in extreme observations, which can be important in some situations.

EXAMPLE 2.18

The median and IQR did not change under the three scenarios in Figure 2.12. 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 stable in the three data sets, the median and IQR estimates are also stable.

[11] These visual estimates will vary a little from one person to the next: $Q_1 = 8\%$, $Q_3 = 14\%$, IQR $= Q_3 - Q_1 = 6\%$. (The true values: $Q_1 = 7.96\%$, $Q_3 = 13.72\%$, IQR $= 5.76\%$.)

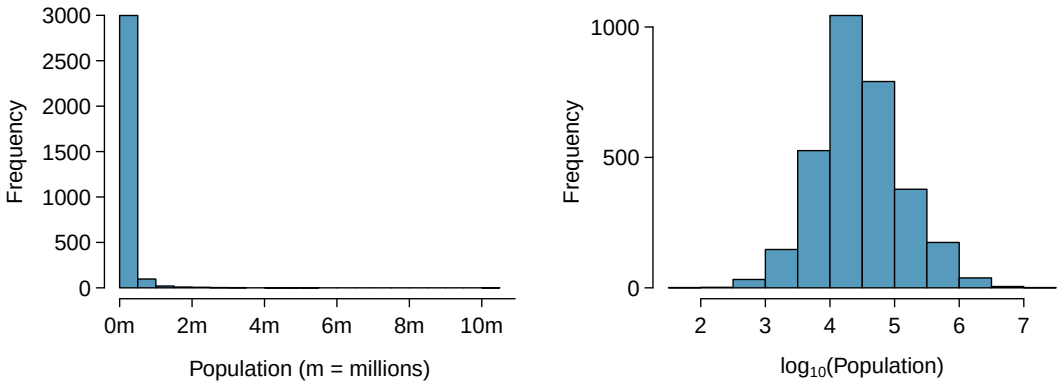
[12] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 2.17.

GUIDED PRACTICE 2.19

G The distribution of loan amounts in the `loan50` data set is right skewed, with a few large loans lingering out into the right tail. If you were wanting to understand the typical loan size, should you be more interested in the mean or median?[13]

2.1.7 Transforming data (special topic)

When data are very strongly skewed, we sometimes transform them so they are easier to model.



(a) (b)

Figure 2.13: (a) A histogram of the populations of all US counties. (b) A histogram of $\log_{10}$-transformed county populations. For this plot, the x-value corresponds to the power of 10, e.g. "4" on the x-axis corresponds to $10^4 = 10{,}000$.

EXAMPLE 2.20

E Consider the histogram of county populations shown in Figure 2.13(a), which shows extreme skew. What isn't useful about this plot?

Nearly all of the data fall into the left-most bin, and the extreme skew obscures many of the potentially interesting details in the data.

There are some standard transformations that may be useful for strongly right skewed data where much of the data is positive but clustered near zero. A **transformation** is a rescaling of the data using a function. For instance, a plot of the logarithm (base 10) of county populations results in the new histogram in Figure 2.13(b). This data is symmetric, and any potential outliers appear much less extreme than in the original data set. By reigning in the outliers and extreme skew, transformations like this often make it easier to build statistical models against the data.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the the population change from 2010 to 2017 against the population in 2010 is shown in Figure 2.14(a). In this first scatterplot, it's hard to decipher any interesting patterns because the population variable is so strongly skewed. However, if we apply a $\log_{10}$ transformation to the population variable, as shown in Figure 2.14(b), a positive association between the variables is revealed. In fact, we may be interested in fitting a trend line to the data when we explore methods around fitting regression lines in Chapter 8.

Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are commonly used by data scientists. Com-

[13] Answers will vary! If we're looking to simply understand what a typical individual loan looks like, the median is probably more useful. However, if the goal is to understand something that scales well, such as the total amount of money we might need to have on hand if we were to offer 1,000 loans, then the mean would be more useful.

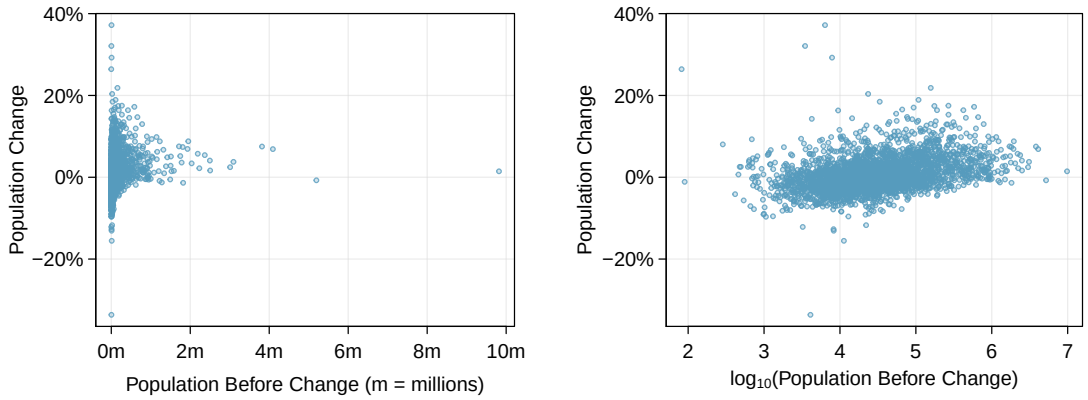


Figure 2.14: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

mon goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

2.1.8 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 create an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 2.15 and 2.16 shows intensity maps for poverty rate in percent (`poverty`), unemployment rate (`unemployment_rate`), homeownership rate in percent (`homeownership`), and median household income (`median_hh_income`). The color key indicates which colors correspond to which values. The intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions or hypotheses.

EXAMPLE 2.21

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

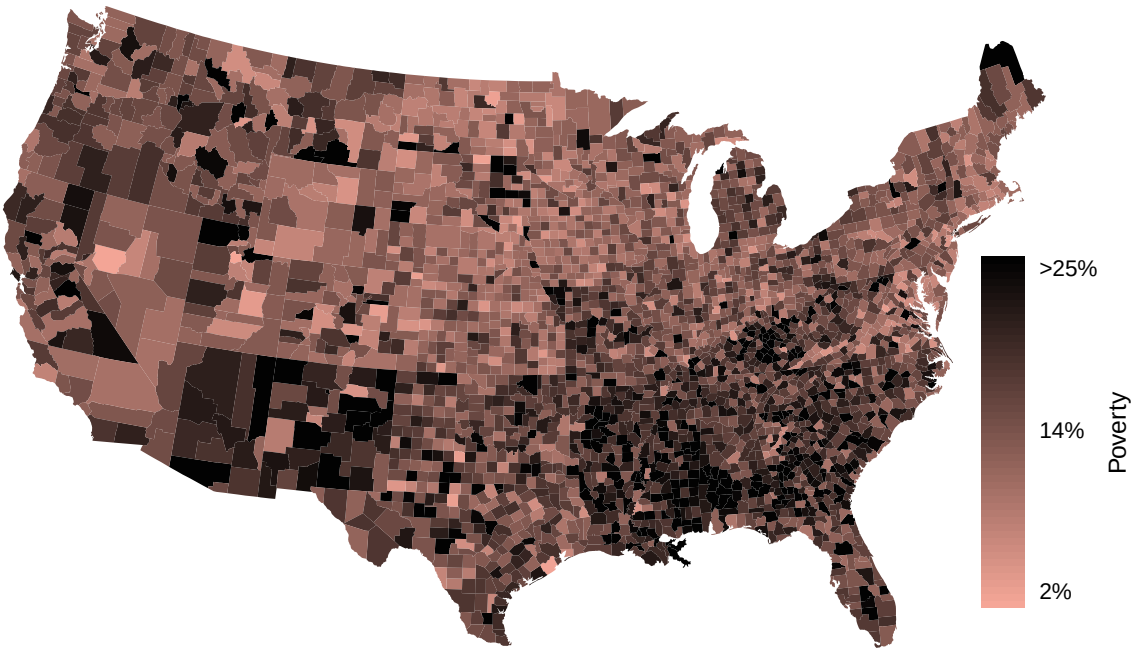
Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does much of Arizona and New Mexico. High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky.

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

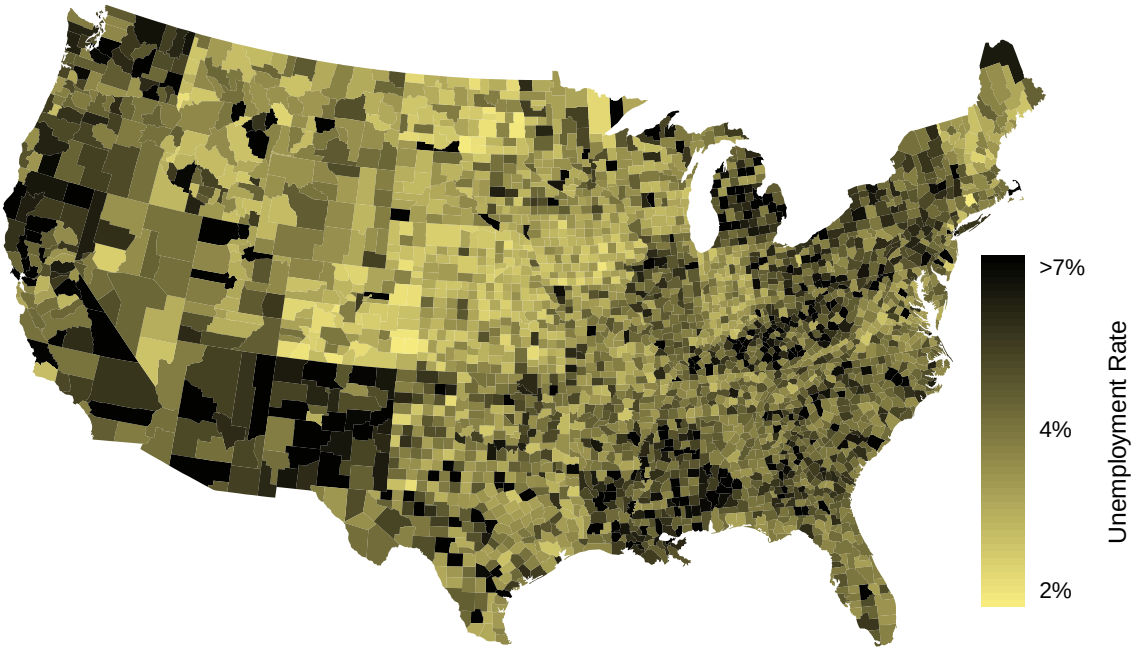
GUIDED PRACTICE 2.22

What interesting features are evident in the `median_hh_income` intensity map in Figure 2.16(b)?[14]

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

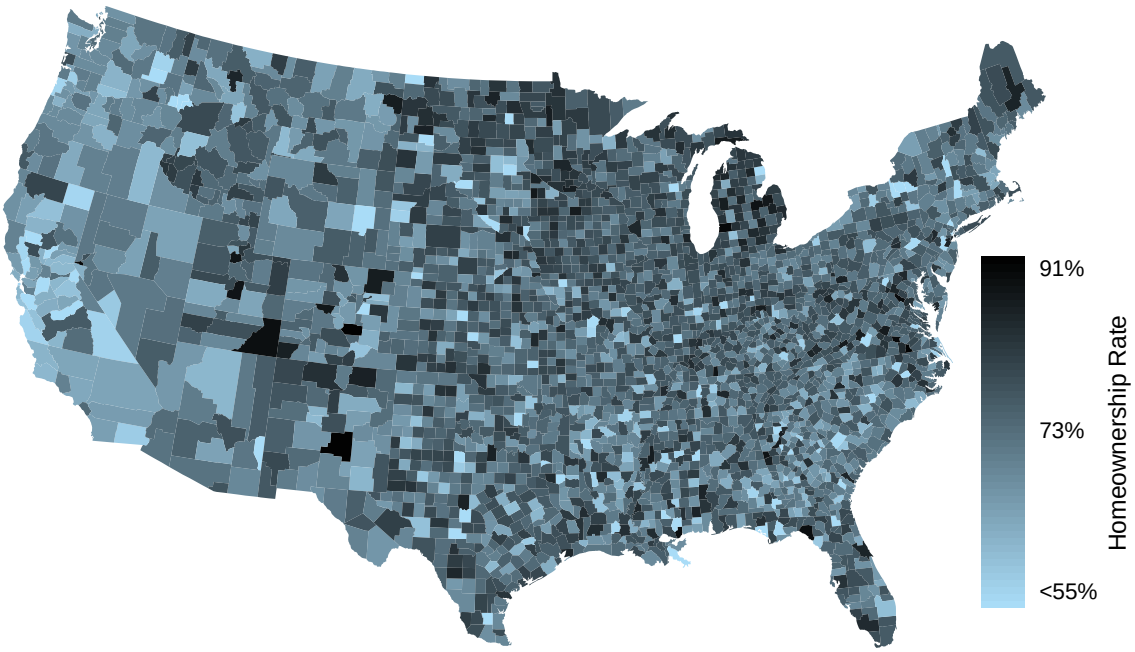


(a)

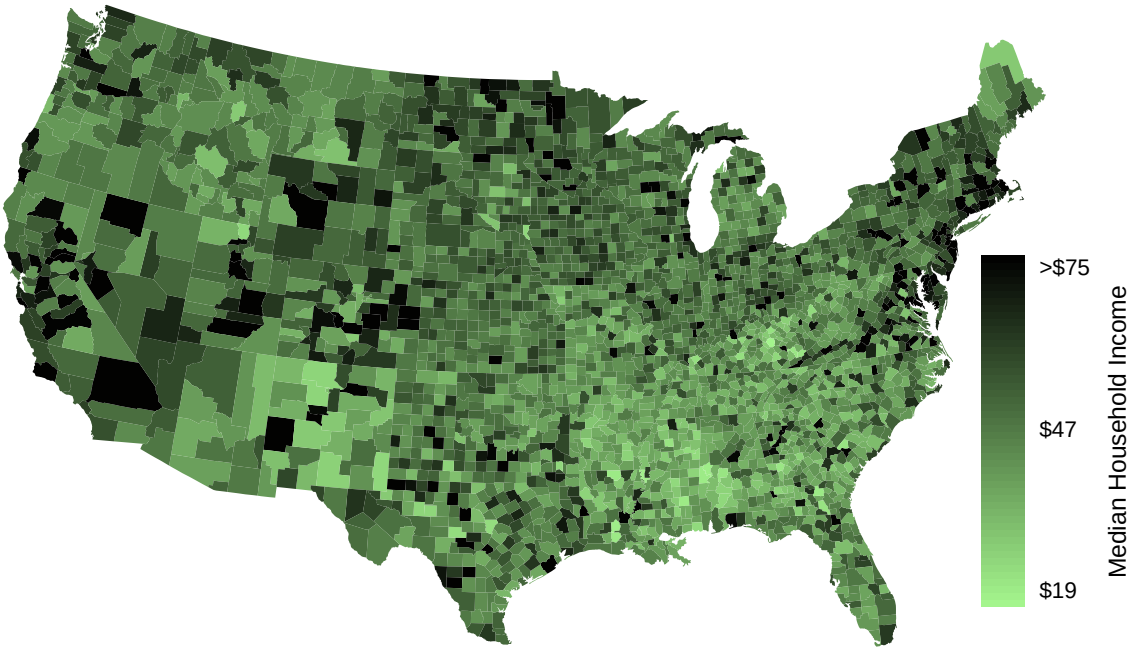


(b)

Figure 2.15: (a) Intensity map of poverty rate (percent). (b) Map of the unemployment rate (percent).



(a)



(b)

Figure 2.16: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income ($1000s).

2.2 Considering categorical data

In this section, we will introduce tables and other basic tools for categorical data that are used throughout this book. The `loan50` data set represents a sample from a larger loan data set called `loans`. This larger data set contains information on 10,000 loans made through Lending Club. We will examine the relationship between `homeownership`, which for the `loans` data can take a value of `rent`, `mortgage` (owns but has a mortgage), or `own`, and `app_type`, which indicates whether the loan application was made with a partner or whether it was an individual application.

2.2.1 Contingency tables and bar plots

Figure 2.17 summarizes two variables: `app_type` and `homeownership`. A table that summarizes data for two categorical variables in this way is called a **contingency table**. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 3496 corresponds to the number of loans in the data set where the borrower rents their home and the application type was by an individual. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $3496 + 3839 + 1170 = 8505$), and **column totals** are total counts down each column. We can also create a table that shows only the overall percentages or proportions for each combination of categories, or we can create a table for a single variable, such as the one shown in Figure 2.18 for the `homeownership` variable.

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Figure 2.17: A contingency table for `app_type` and `homeownership`.

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Figure 2.18: A table summarizing the frequencies of each value for the `homeownership` variable.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 2.19 shows a **bar plot** for the `homeownership` variable. In the right panel, the counts are converted into proportions, showing the proportion of observations that are in each level (e.g. $3858/10000 = 0.3858$ for `rent`).

2.2.2 Row and column proportions

Sometimes it is useful to understand the fractional breakdown of one variable in another, and we can modify our contingency table to provide such a view. Figure 2.20 shows the **row proportions** for Figure 2.17, which are computed as the counts divided by their row totals. The value 3496 at the intersection of `individual` and `rent` is replaced by $3496/8505 = 0.411$, i.e. 3496 divided by its row total, 8505. So what does 0.411 represent? It corresponds to the proportion of individual applicants who rent.

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Figure 2.21 shows such a table, and here the value 0.906 indicates that 90.6% of renters applied as individuals for the

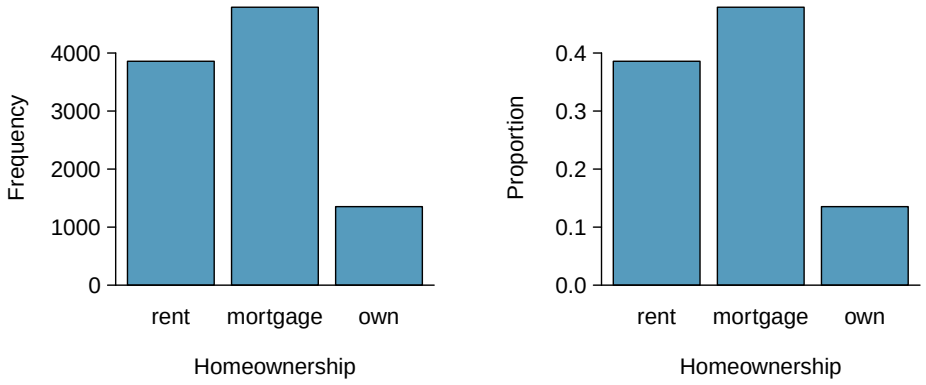


Figure 2.19: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

	rent	mortgage	own	Total
individual	0.411	0.451	0.138	1.000
joint	0.242	0.635	0.122	1.000
Total	0.386	0.479	0.135	1.000

Figure 2.20: A contingency table with row proportions for the `app_type` and `homeownership` variables. The row total is off by 0.001 for the `joint` row due to a rounding error.

loan. This rate is higher compared to loans from people with mortgages (80.2%) or who own their home (85.1%). Because these rates vary between the three levels of `homeownership` (`rent`, `mortgage`, `own`), this provides evidence that the `app_type` and `homeownership` variables are associated.

	rent	mortgage	own	Total
individual	0.906	0.802	0.865	0.851
joint	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

Figure 2.21: A contingency table with column proportions for the `app_type` and `homeownership` variables. The total for the last column is off by 0.001 due to a rounding error.

We could also have checked for an association between `app_type` and `homeownership` in Figure 2.20 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of loans where the borrower rents, has a mortgage, or owns varied across the `individual` to `joint` application types.

GUIDED PRACTICE 2.23

(a) What does 0.451 represent in Figure 2.20?
(b) What does 0.802 represent in Figure 2.21?[15]

GUIDED PRACTICE 2.24

(a) What does 0.122 at the intersection of `joint` and `own` represent in Figure 2.20?
(b) What does 0.135 represent in the Figure 2.21?[16]

[15] (a) 0.451 represents the proportion of individual applicants who have a mortgage. (b) 0.802 represents the fraction of applicants with mortgages who applied as individuals.

[16] (a) 0.122 represents the fraction of joint borrowers who own their home. (b) 0.135 represents the home-owning borrowers who had a joint application for the loan.

EXAMPLE 2.25

Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One such characteristic is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is the email format, which indicates whether or not an email has any HTML content, such as bolded text. We'll focus on email format and spam status using the `email` data set, and these variables are summarized in a contingency table in Figure 2.22. Which would be more helpful to someone hoping to classify email as spam or regular email for this table: row or column proportions?

(E)

A data scientist would be interested in how the proportion of spam changes within each email format. This corresponds to column proportions: the proportion of spam in plain text emails and the proportion of spam in HTML emails.

If we generate the column proportions, we can see that a higher fraction of plain text emails are spam ($209/1195 = 17.5\%$) than compared to HTML emails ($158/2726 = 5.8\%$). This information on its own is insufficient to classify an email as spam or not spam, as over 80% of plain text emails are not spam. Yet, when we carefully combine this information with many other characteristics, we stand a reasonable chance of being able to classify some emails as spam or not spam with confidence.

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

Figure 2.22: A contingency table for `spam` and `format`.

Example 2.25 points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed. However, sometimes it simply isn't clear which, if either, is more useful.

EXAMPLE 2.26

Look back to Tables 2.20 and 2.21. Are there any obvious scenarios where one might be more useful than the other?

(E)

None that we thought were obvious! What is distinct about `app_type` and `homeownership` vs the email example is that these two variables don't have a clear explanatory-response variable relationship that we might hypothesize (see Section 1.2.4 for these terms). Usually it is most useful to "condition" on the explanatory variable. For instance, in the email example, the email format was seen as a possible explanatory variable of whether the message was spam, so we would find it more interesting to compute the relative frequencies (proportions) for each email format.

2.2.3 Using a bar plot with two variables

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

A **stacked bar plot** is a graphical display of contingency table information. For example, a stacked bar plot representing Figure 2.21 is shown in Figure 2.23(a), where we have first created a bar plot using the `homeownership` variable and then divided each group by the levels of `app_type`.

One related visualization to the stacked bar plot is the **side-by-side bar plot**, where an example is shown in Figure 2.23(b).

For the last type of bar plot we introduce, the column proportions for the `app_type` and `homeownership` contingency table have been translated into a standardized stacked bar plot in Figure 2.23(c). This type of visualization is helpful in understanding the fraction of individual

or joint loan applications for borrowers in each level of `homeownership`. Additionally, since the proportions of `joint` and `individual` vary across the groups, we can conclude that the two variables are associated.

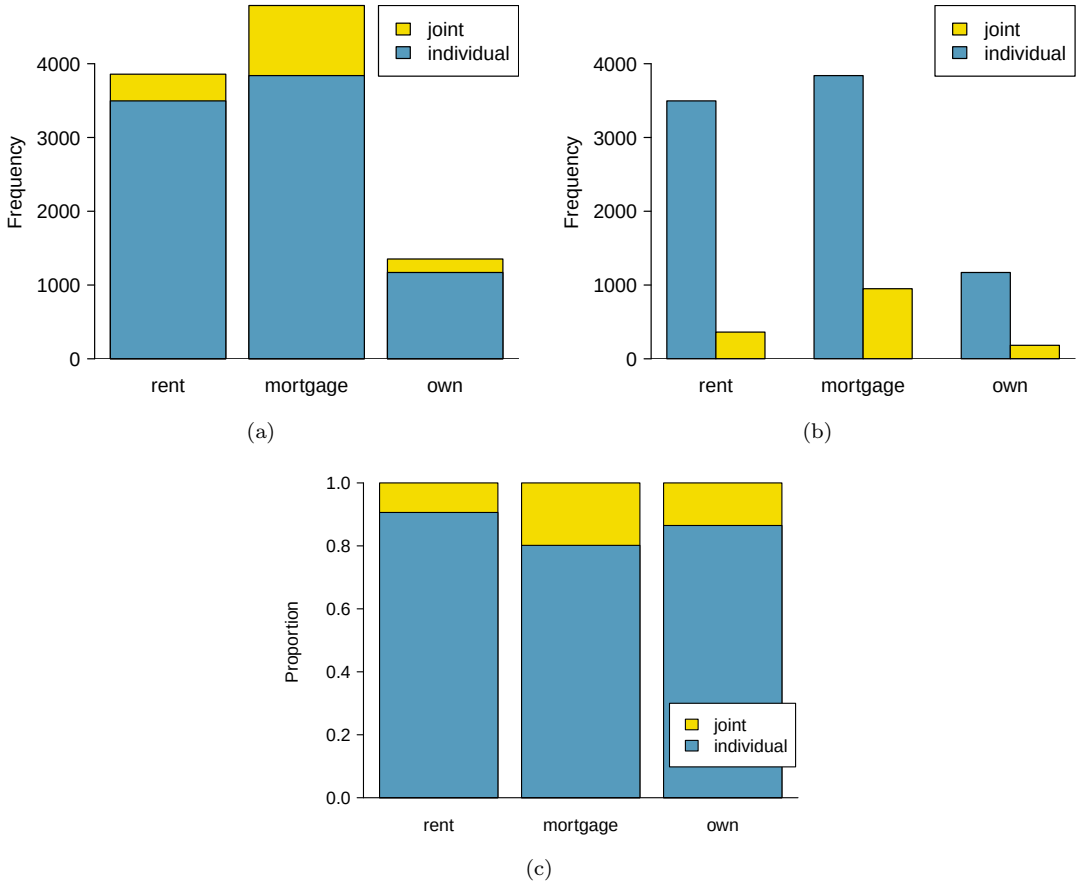


Figure 2.23: (a) Stacked bar plot for `homeownership`, where the counts have been further broken down by `app_type`. (b) Side-by-side bar plot for `homeownership` and `app_type`. (c) Standardized version of the stacked bar plot. **??** A contingency table for `app_type` and `homeownership`.

EXAMPLE 2.27

Examine the three bar plots in Figure 2.23. When is the stacked, side-by-side, or standardized stacked bar plot the most useful?

The stacked bar plot is most useful when it's reasonable to assign one variable as the explanatory variable and the other variable as the response, since we are effectively grouping by one variable first and then breaking it down by the others.

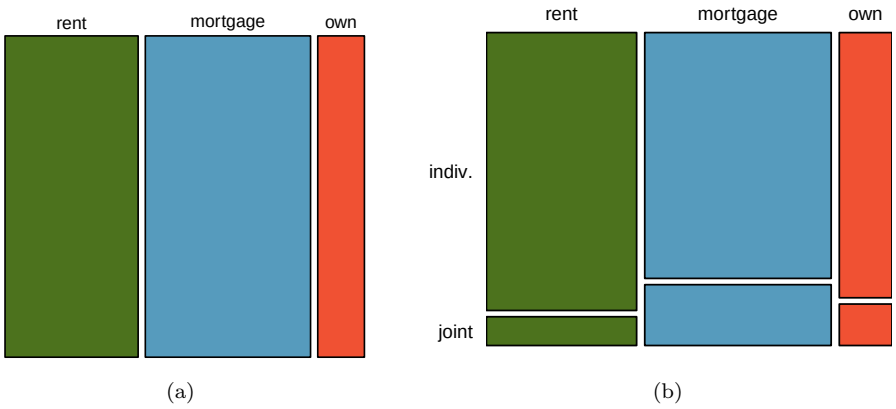
(E) Side-by-side bar plots are more agnostic in their display about which variable, if any, represents the explanatory and which the response variable. It is also easy to discern the number of cases in of the six different group combinations. However, one downside is that it tends to require more horizontal space; the narrowness of Figure 2.23(b) makes the plot feel a bit cramped. Additionally, when two groups are of very different sizes, as we see in the `own` group relative to either of the other two groups, it is difficult to discern if there is an association between the variables.

The standardized stacked bar plot is helpful if the primary variable in the stacked bar plot is relatively imbalanced, e.g. the `own` category has only a third of the observations in the `mortgage` category, making the simple stacked bar plot less useful for checking for an association. The major downside of the standardized version is that we lose all sense of how many cases each of the bars represents.

2.2.4 Mosaic plots

A **mosaic plot** is a visualization technique suitable for contingency tables that resembles a standardized stacked bar plot with the benefit that we still see the relative group sizes of the primary variable as well.

To get started in creating our first mosaic plot, we'll break a square into columns for each category of the `homeownership` variable, with the result shown in Figure 2.24(a). Each column represents a level of `homeownership`, and the column widths correspond to the proportion of loans in each of those categories. For instance, there are fewer loans where the borrower is an owner than where the borrower has a mortgage. In general, mosaic plots use box *areas* to represent the number of cases in each category.



(a) (b)

Figure 2.24: (a) The one-variable mosaic plot for `homeownership`. (b) Two-variable mosaic plot for both `homeownership` and `app_type`.

To create a completed mosaic plot, the single-variable mosaic plot is further divided into pieces in Figure 2.24(b) using the `app_type` variable. Each column is split proportional to the number of loans from individual and joint borrowers. For example, the second column represents loans where the borrower has a mortgage, and it was divided into individual loans (upper) and joint loans (lower). As another example, the bottom segment of the third column represents loans where the borrower owns their home and applied jointly, while the upper segment of this column represents borrowers who are homeowners and filed individually. We can again use this plot to see that the `homeownership` and `app_type` variables are associated, since some columns are divided in different vertical locations than others, which was the same technique used for checking an association in the standardized stacked bar plot.

In Figure 2.25, we chose to first split by the homeowner status of the borrower. However, we could have instead first split by the application type, as in Figure 2.25. Like with the bar plots, it's common to use the explanatory variable to represent the first split in a mosaic plot, and then for the response to break up each level of the explanatory variable, if these labels are reasonable to attach to the variables under consideration.

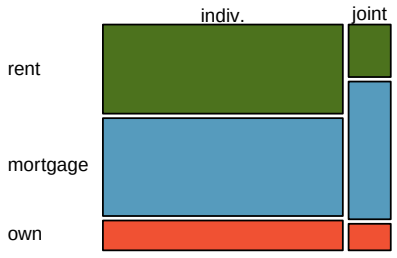


Figure 2.25: Mosaic plot where loans are grouped by the `homeownership` variable after they've been divided into the `individual` and `joint` application types.

2.2.5 The only pie chart you will see in this book

A **pie chart** is shown in Figure 2.26 alongside a bar plot representing the same information. Pie charts can be useful for giving a high-level overview to show how a set of cases break down. However, it is also difficult to decipher details in a pie chart. For example, it takes a couple seconds longer to recognize that there are more loans where the borrower has a mortgage than rent when looking at the pie chart, while this detail is very obvious in the bar plot. While pie charts can be useful, we prefer bar plots for their ease in comparing groups.

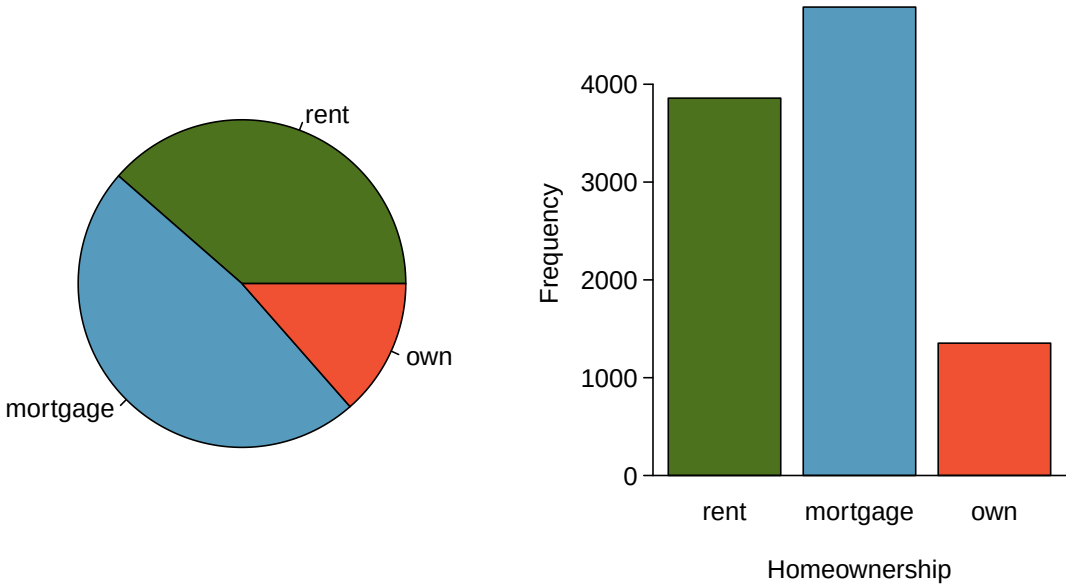


Figure 2.26: A pie chart and bar plot of `homeownership`.

2.2.6 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's required is to make a numerical plot for each group in the same graph. 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 2010 to 2017 versus counties that had no gain. While we might like to make a causal connection here, remember that these are observational data and so such an interpretation would be, at best, half-baked.

There were 1,454 counties where the population increased from 2010 to 2017, and there were 1,672 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Figure 2.27 to give a better sense of some of the raw median income data.

The **side-by-side box plot** is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 2.28, where there are two box plots, one for each group, placed into one plotting window and drawn on the same scale.

Another useful plotting method uses **hollow histograms** to compare numerical data across groups. These are just the outlines of histograms of each group put on the same plot, as shown in the right panel of Figure 2.28.

Median Income for 150 Counties, in $1000s

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

Figure 2.27: In this table, median household income (in $1000s) from a random sample of 100 counties that had population gains are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

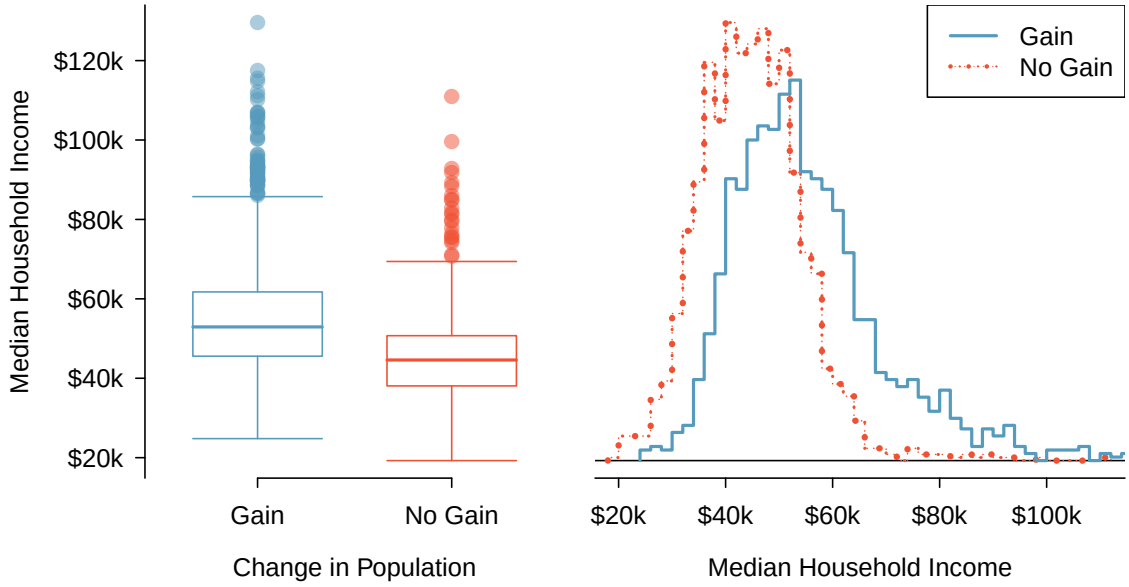


Figure 2.28: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_hh_income`, where the counties are split by whether there was a population gain or loss.

GUIDED PRACTICE 2.28

Use the plots in Figure 2.28 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?[17]

GUIDED PRACTICE 2.29

What components of each plot in Figure 2.28 do you find most useful?[18]

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

[18] 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 potential anomalies.

2.3 Case study: malaria vaccine

EXAMPLE 2.30

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 $\hat{p}_R$?

(E)

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

GUIDED PRACTICE 2.31

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?[19]

(G)

2.3.1 Variability within data

We consider a study on a new malaria vaccine called PfSPZ. In this study, volunteer patients were randomized into one of two experiment groups: 14 patients received an experimental vaccine or 6 patients received a placebo vaccine. Nineteen weeks later, all 20 patients were exposed to a drug-sensitive malaria virus strain; the motivation of using a drug-sensitive strain of virus here is for ethical considerations, allowing any infections to be treated effectively. The results are summarized in Figure 2.29, where 9 of the 14 treatment patients remained free of signs of infection while all of the 6 patients in the control group patients showed some baseline signs of infection.

		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figure 2.29: Summary results for the malaria vaccine experiment.

GUIDED PRACTICE 2.32

Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[20]

(G)

In this study, a smaller proportion of patients who received the vaccine showed signs of an infection (35.7% versus 100%). However, the sample is very small, and it is unclear whether the difference provides *convincing evidence* that the vaccine is effective.

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

[20] The study is an experiment, as patients were randomly assigned an experiment group. Since this is an experiment, the results can be used to evaluate a causal relationship between the malaria vaccine and whether patients showed signs of an infection.

EXAMPLE 2.33

Data scientists are sometimes called upon to evaluate the strength of evidence. When looking at the rates of infection for patients in the two groups in this study, what comes to mind as we try to determine whether the data show convincing evidence of a real difference?

(E) The observed infection rates (35.7% for the treatment group versus 100% for the control group) suggest the vaccine may be effective. However, we cannot be sure if the observed difference represents the vaccine's efficacy or is just from random chance. Generally there is a little bit of fluctuation in sample data, and we wouldn't expect the sample proportions to be *exactly* equal, even if the truth was that the infection rates were independent of getting the vaccine. Additionally, with such small samples, perhaps it's common to observe such large differences when we randomly split a group due to chance alone!

Example 2.33 is a reminder that the observed outcomes in the data sample may not perfectly reflect the true relationships between variables since there is **random noise**. While the observed difference in rates of infection is large, the sample size for the study is small, making it unclear if this observed difference represents efficacy of the vaccine or whether it is simply due to chance. We label these two competing claims, H_0 and H_A, which are spoken as "H-nought" and "H-A":

H_0: **Independence model.** The variables `treatment` and `outcome` are independent. They have no relationship, and the observed difference between the proportion of patients who developed an infection in the two groups, 64.3%, was due to chance.

H_A: **Alternative model.** The variables are *not* independent. The difference in infection rates of 64.3% was not due to chance, and vaccine affected the rate of infection.

What would it mean if the independence model, which says the vaccine had no influence on the rate of infection, is true? It would mean 11 patients were going to develop an infection *no matter which group they were randomized into*, and 9 patients would not develop an infection *no matter which group they were randomized into*. That is, if the vaccine did not affect the rate of infection, the difference in the infection rates was due to chance alone in how the patients were randomized.

Now consider the alternative model: infection rates were influenced by whether a patient received the vaccine or not. If this was true, and especially if this influence was substantial, we would expect to see some difference in the infection rates of patients in the groups.

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

2.3.2 Simulating the study

We're going to implement **simulations**, where we will pretend we know that the malaria vaccine being tested does *not* work. Ultimately, we want to understand if the large difference we observed is common in these simulations. If it is common, then maybe the difference we observed was purely due to chance. If it is very uncommon, then the possibility that the vaccine was helpful seems more plausible.

Figure 2.29 shows that 11 patients developed infections and 9 did not. For our simulation, we will suppose the infections were independent of the vaccine and we were able to *rewind* back to when the researchers randomized the patients in the study. If we happened to randomize the patients differently, we may get a different result in this hypothetical world where the vaccine doesn't influence the infection. Let's complete another **randomization** using a simulation.

In this **simulation**, we take 20 notecards to represent the 20 patients, where we write down "infection" on 11 cards and "no infection" on 9 cards. In this hypothetical world, we believe each patient that got an infection was going to get it regardless of which group they were in, so let's see what happens if we randomly assign the patients to the treatment and control groups again. We thoroughly shuffle the notecards and deal 14 into a `vaccine` pile and 6 into a `placebo` pile. Finally, we tabulate the results, which are shown in Figure 2.30.

		outcome		
		infection	no infection	Total
treatment (simulated)	vaccine	7	7	14
	placebo	4	2	6
	Total	11	9	20

Figure 2.30: Simulation results, where any difference in infection rates is purely due to chance.

GUIDED PRACTICE 2.34

G What is the difference in infection rates between the two simulated groups in Figure 2.30? How does this compare to the observed 64.3% difference in the actual data?[21]

2.3.3 Checking for independence

We computed one possible difference under the independence model in Guided Practice 2.34, which represents one difference due to chance. While in this first simulation, we physically dealt out notecards to represent the patients, it is more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance:

$$\frac{2}{6} - \frac{9}{14} = -0.310$$

And another:

$$\frac{3}{6} - \frac{8}{14} = -0.071$$

And so on until we repeat the simulation enough times that we have a good idea of what represents the *distribution of differences from chance alone*. Figure 2.31 shows a stacked plot of the differences found from 100 simulations, where each dot represents a simulated difference between the infection rates (control rate minus treatment rate).

Note that the distribution of these simulated differences is centered around 0. We simulated these differences assuming that the independence model was true, and under this condition, we expect the difference to be near zero with some random fluctuation, where *near* is pretty generous in this case since the sample sizes are so small in this study.

EXAMPLE 2.35

How often would you observe a difference of at least 64.3% (0.643) according to Figure 2.31? Often, sometimes, rarely, or never?

E

It appears that a difference of at least 64.3% due to chance alone would only happen about 2% of the time according to Figure 2.31. Such a low probability indicates a rare event.

The difference of 64.3% being a rare event suggests two possible interpretations of the results of the study:

H_0 **Independence model.** The vaccine has no effect on infection rate, and we just happened to observe a difference that would only occur on a rare occasion.

H_A **Alternative model.** The vaccine has an effect on infection rate, and the difference we observed was actually due to the vaccine being effective at combatting malaria, which explains the large difference of 64.3%.

Based on the simulations, we have two options. (1) We conclude that the study results do not

[21] $4/6 - 7/14 = 0.167$ or about 16.7% in favor of the vaccine. This difference due to chance is much smaller than the difference observed in the actual groups.

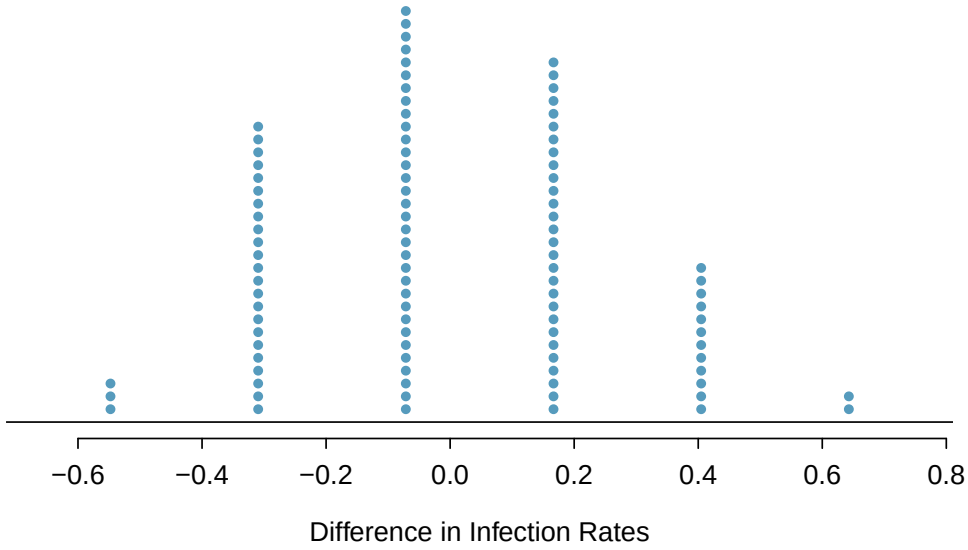


Figure 2.31: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

provide strong evidence against the independence model. That is, we do not have sufficiently strong evidence to conclude the vaccine had an effect in this clinical setting. (2) We conclude the evidence is sufficiently strong to reject H_0 and assert that the vaccine was useful. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event.[22] So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence that the vaccine provides some protection against malaria in this clinical setting.

One field of statistics, statistical inference, is built on evaluating whether such differences are due to chance. In statistical inference, data scientists evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur. In Chapter 5, we give a formal introduction to the problem of model selection. We spend the next two chapters building a foundation of probability and theory necessary to make that discussion rigorous.

[22] 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 292 million chance that the Powerball numbers for the largest jackpot in history (January 13th, 2016) would be (04, 08, 19, 27, 34) with a Powerball of (10), 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 3

Probability

Probability forms a foundation for statistics. and you're probably already familiar with many of the ideas. However, formalization of the concepts is new for most. This chapter aims to introduce probability concepts using many examples that will be familiar to most people.

While this chapter provides a theoretical foundation for the ideas introduced in later chapters, mastery of all concepts introduced in this chapter is not strictly required to understand or apply the methods introduced in the rest of this book.

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

3.1 Defining probability

Statistics is based on probability, and while probability is not required for the applied techniques in this book, it may help you gain a deeper understanding of the methods and set a better foundation for future courses.

3.1.1 Introductory examples

Before we get into technical ideas, let's walk through some basic examples that may feel more familiar.

EXAMPLE 3.1

A "die", the singular of dice, is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting 1 when rolling a die?

(E)

If the die is fair, then the chance of a 1 is as good as the chance of any other number. Since there are six outcomes, the chance must be 1-in-6 or, equivalently, 1/6.

EXAMPLE 3.2

What is the chance of getting a 1 or 2 in the next roll?

(E)

1 and 2 constitute two of the six equally likely possible outcomes, so the chance of getting one of these two outcomes must be $2/6 = 1/3$.

EXAMPLE 3.3

What is the chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?

(E)

100%. The outcome must be one of these numbers.

EXAMPLE 3.4

What is the chance of not rolling a 2?

(E)

Since the chance of rolling a 2 is 1/6 or $16.\bar{6}\%$, the chance of not rolling a 2 must be $100\% - 16.\bar{6}\% = 83.\bar{3}\%$ or 5/6.

Alternatively, we could have noticed that not rolling a 2 is the same as getting a 1, 3, 4, 5, or 6, which makes up five of the six equally likely outcomes and has probability 5/6.

EXAMPLE 3.5

Consider rolling two dice. If 1/6 of the time the first die is a 1 and 1/6 of those times the second die is a 1, what is the chance of getting two 1s?

(E)

If $16.\bar{6}\%$ of the time the first die is a 1 and 1/6 of *those* times the second die is also a 1, then the chance that both dice are 1 is $(1/6) \times (1/6)$ or 1/36.

3.1.2 Probability

We use probability to build tools to describe and understand apparent randomness. We often frame probability in terms of a **random process** giving rise to an **outcome**.

Roll a die	$\rightarrow$	1, 2, 3, 4, 5, or 6
Flip a coin	$\rightarrow$	H or T

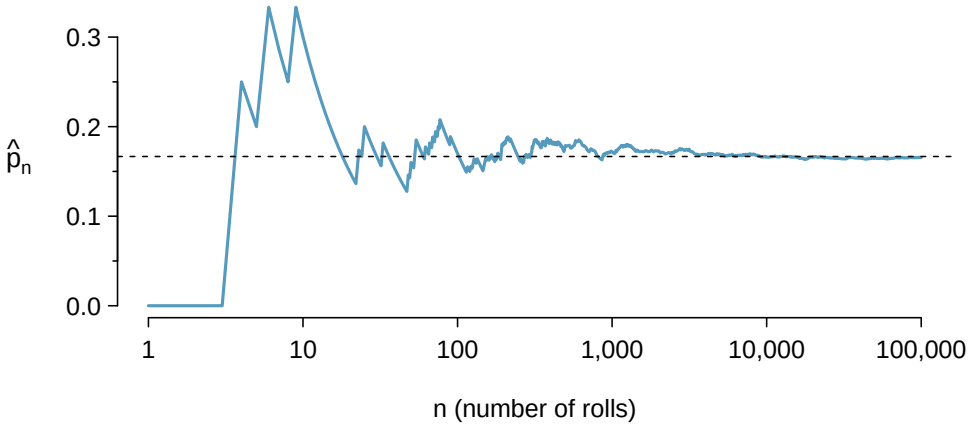


Figure 3.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

Rolling a die or flipping a coin is a seemingly random process and each gives rise to an outcome.

PROBABILITY

The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be displayed as a percentage between 0% and 100%.

Probability can be illustrated by rolling a die many times. Let $\hat{p}_n$ be the proportion of outcomes that are `1` after the first n rolls. As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a `1`, $p = 1/6$. Figure 3.1 shows this convergence for 100,000 die rolls. The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**.

LAW OF LARGE NUMBERS

As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

Occasionally the proportion will veer off from the probability and appear to defy the Law of Large Numbers, as $\hat{p}_n$ does many times in Figure 3.1. However, these deviations become smaller as the number of rolls increases.

Above we write p as the probability of rolling a `1`. We can also write this probability as

$$P(\text{rolling a } \texttt{1})$$

As we become more comfortable with this notation, we will abbreviate it further. For instance, if it is clear that the process is "rolling a die", we could abbreviate $P(\text{rolling a } \texttt{1})$ as $P(\texttt{1})$.

GUIDED PRACTICE 3.6

(G) Random processes include rolling a die and flipping a coin. (a) Think of another random process. (b) Describe all the possible outcomes of that process. For instance, rolling a die is a random process with possible outcomes `1`, `2`, ..., `6`.[1]

[1] Here are four examples. (i) Whether someone gets sick in the next month or not is an apparently random process with outcomes `sick` and `not`. (ii) We can *generate* a random process by randomly picking a person and measuring that person's height. The outcome of this process will be a positive number. (iii) Whether the stock market goes up or down next week is a seemingly random process with possible outcomes `up`, `down`, and `no_change`. Alternatively, we

What we think of as random processes are not necessarily random, but they may just be too difficult to understand exactly. The fourth example in the footnote solution to Guided Practice 3.6 suggests a roommate's behavior is a random process. However, even if a roommate's behavior is not truly random, modeling her behavior as a random process can still be useful.

3.1.3 Disjoint or mutually exclusive outcomes

Two outcomes are called **disjoint** or **mutually exclusive** if they cannot both happen. For instance, if we roll a die, the outcomes `1` and `2` are disjoint since they cannot both occur. On the other hand, the outcomes `1` and "rolling an odd number" are not disjoint since both occur if the outcome of the roll is a `1`. The terms *disjoint* and *mutually exclusive* are equivalent and interchangeable.

Calculating the probability of disjoint outcomes is easy. When rolling a die, the outcomes `1` and `2` are disjoint, and we compute the probability that one of these outcomes will occur by adding their separate probabilities:

$$P(\texttt{1 or 2}) = P(\texttt{1}) + P(\texttt{2}) = 1/6 + 1/6 = 1/3$$

What about the probability of rolling a `1`, `2`, `3`, `4`, `5`, or `6`? Here again, all of the outcomes are disjoint so we add the probabilities:

$$\begin{aligned} &P(\texttt{1 or 2 or 3 or 4 or 5 or 6}) \\ &\quad = P(\texttt{1}) + P(\texttt{2}) + P(\texttt{3}) + P(\texttt{4}) + P(\texttt{5}) + P(\texttt{6}) \\ &\quad = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1 \end{aligned}$$

The **Addition Rule** guarantees the accuracy of this approach when the outcomes are disjoint.

ADDITION RULE OF DISJOINT OUTCOMES

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k)$$

GUIDED PRACTICE 3.7

(G) We are interested in the probability of rolling a `1`, `4`, or `5`. (a) Explain why the outcomes `1`, `4`, and `5` are disjoint. (b) Apply the Addition Rule for disjoint outcomes to determine $P(\texttt{1 or 4 or 5})$.[2]

[The next Guided Practice is new, replacing one that had used the `email` data set.]

GUIDED PRACTICE 3.8

(G) In the `loans` data set in Chapter 2, the `homeownership` variable described whether the borrower rents, has a mortgage, or owns her property. Of the 10,000 borrowers, 3858 rented, 4789 had a mortgage, and 1353 owned her home. (a) Are the outcomes `rent`, `mortgage`, and `own` disjoint? (b) Determine the proportion of loans with value `mortgage` and `own` separately. (c) Use the Addition Rule for disjoint outcomes to compute the probability a randomly selected loan from the data set is for someone who has a mortgage or owns her home.[3]

could have used the percent change in the stock market as a numerical outcome. (iv) Whether your roommate cleans her dishes tonight probably seems like a random process with possible outcomes `cleans_dishes` and `leaves_dishes`.

[2](a) The random process is a die roll, and at most one of these outcomes can come up. This means they are disjoint outcomes. (b) $P(\texttt{1 or 4 or 5}) = P(\texttt{1}) + P(\texttt{4}) + P(\texttt{5}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

Data scientists rarely work with individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where a die roll results in `1` or `2` and B represent the event that the die roll is a `4` or a `6`. We write A as the set of outcomes $\{1, 2\}$ and $B = \{4, 6\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 3.2.

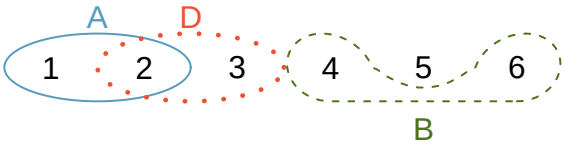


Figure 3.2: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

The Addition Rule applies to both disjoint outcomes and disjoint events. The probability that one of the disjoint events A or B occurs is the sum of the separate probabilities:

$$P(A \text{ or } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

GUIDED PRACTICE 3.9

(a) Verify the probability of event A, $P(A)$, is 1/3 using the Addition Rule. (b) Do the same for event B.[4]

GUIDED PRACTICE 3.10

(a) Using Figure 3.2 as a reference, what outcomes are represented by event D? (b) Are events B and D disjoint? (c) Are events A and D disjoint?[5]

GUIDED PRACTICE 3.11

In Guided Practice 3.10, you confirmed B and D from Figure 3.2 are disjoint. Compute the probability that event B or event D occurs.[6]

3.1.4 Probabilities when events are not disjoint

Let's consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Figure 3.3. If you are unfamiliar with the cards in a regular deck, please see the footnote.[7]

GUIDED PRACTICE 3.12

(a) What is the probability that a randomly selected card is a diamond? (b) What is the probability that a randomly selected card is a face card?[8]

[3](a) Yes. Each loan is categorized in only one level of `homeownership`. (b) Mortgage: $\frac{4789}{10000} = 0.479$. Own: $\frac{1353}{10000} = 0.135$. (c) $P(\texttt{mortgage or own}) = P(\texttt{mortgage}) + P(\texttt{own}) = 0.479 + 0.135 = 0.614$.

[4](a) $P(A) = P(\texttt{1 or 2}) = P(\texttt{1}) + P(\texttt{2}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Similarly, $P(B) = 1/3$.

[5](a) Outcomes `2` and `3`. (b) Yes, events B and D are disjoint because they share no outcomes. (c) The events A and D share an outcome in common, `2`, and so are not disjoint.

[6]Since B and D are disjoint events, use the Addition Rule: $P(B \text{ or } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

[7]The 52 cards are split into four **suits**: ♣ (club), ♢ (diamond), ♡ (heart), ♠ (spade). Each suit has its 13 cards labeled: `2`, `3`, ..., `10`, `J` (jack), `Q` (queen), `K` (king), and `A` (ace). Thus, each card is a unique combination of a suit and a label, e.g. 4♡ and J♣. The 12 cards represented by the jacks, queens, and kings are called `face cards`. The cards that are ♢ or ♡ are typically colored red while the other two suits are typically colored black.

2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2♢	3♢	4♢	5♢	6♢	7♢	8♢	9♢	10♢	J♢	Q♢	K♢	A♢
2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡	A♡
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

Figure 3.3: Representations of the 52 unique cards in a deck.

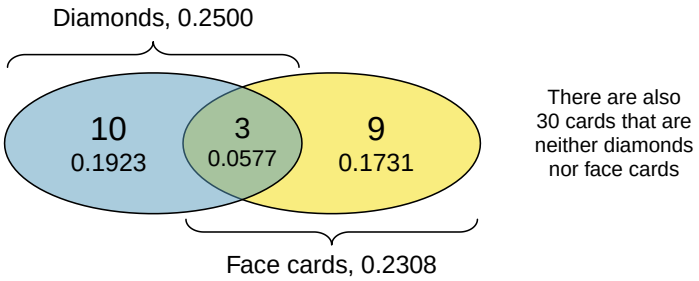


Figure 3.4: A Venn diagram for diamonds and face cards.

Venn diagrams are useful when outcomes can be categorized as "in" or "out" for two or three variables, attributes, or random processes. The Venn diagram in Figure 3.4 uses a circle to represent diamonds and another to represent face cards. If a card is both a diamond and a face card, it falls into the intersection of the circles. If it is a diamond but not a face card, it will be in part of the left circle that is not in the right circle (and so on). The total number of cards that are diamonds is given by the total number of cards in the diamonds circle: $10 + 3 = 13$. The probabilities are also shown (e.g. $10/52 = 0.1923$).

Let A represent the event that a randomly selected card is a diamond and B represent the event that it is a face card. How do we compute $P(A \text{ or } B)$? Events A and B are not disjoint – the cards $J\diamondsuit$, $Q\diamondsuit$, and $K\diamondsuit$ fall into both categories – so we cannot use the Addition Rule for disjoint events. Instead we use the Venn diagram. We start by adding the probabilities of the two events:

$$P(A) + P(B) = P(\diamondsuit) + P(\text{face card}) = 13/52 + 12/52$$

However, the three cards that are in both events were counted twice, once in each probability. We must correct this double counting:

$$\begin{aligned} P(A \text{ or } B) &= P(\diamondsuit \text{ or face card}) \\ &= P(\diamondsuit) + P(\text{face card}) - P(\diamondsuit \text{ and face card}) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned}$$

This equation is an example of the **General Addition Rule**.

GENERAL ADDITION RULE

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where $P(A \text{ and } B)$ is the probability that both events occur.

[8] (a) There are 52 cards and 13 diamonds. If the cards are thoroughly shuffled, each card has an equal chance of being drawn, so the probability that a randomly selected card is a diamond is $P(\diamondsuit) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

TIP: "or" is inclusive
When we write "or" in statistics, we mean "and/or" unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur.

GUIDED PRACTICE 3.13

(G) (a) If A and B are disjoint, describe why this implies $P(A \text{ and } B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the simpler Addition Rule for disjoint events if A and B are disjoint.[9]

GUIDED PRACTICE 3.14

(G) In the `loans` data set describing 10,000 loans, 1495 loans were from joint applications (e.g. a couple applied together), 4789 applicants had a mortgage, and 950 had both of these characteristics. Create a Venn diagram for this setup.[10]

GUIDED PRACTICE 3.15

(G) (a) Use your Venn diagram from Guided Practice 3.14 to determine the probability a randomly drawn loan from the `loans` data set is from a joint application where the couple had a mortgage. (b) What is the probability that the loan had either of these attributes?[11]

3.1.5 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Figure 3.5 shows the probability distribution for the sum of two dice.

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Figure 3.5: Probability distribution for the sum of two dice.

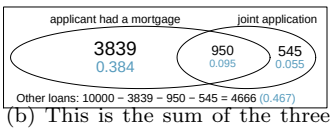
RULES FOR PROBABILITY DISTRIBUTIONS

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

[9] (a) If A and B are disjoint, A and B can never occur simultaneously. (b) If A and B are disjoint, then the last $P(A \text{ and } B)$ term of in the General Addition Rule formula is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

[10] Both the counts and corresponding probabilities (e.g. 3839/10000 = 0.384) are shown. Notice that the number of loans represented in the left circle corresponds to $3839 + 950 = 4789$, and the number represented in the right circle is $950 + 545 = 1495$.



[11] (a) The solution is represented by the intersection of the two circles: 0.095. (b) This is the sum of the three disjoint probabilities shown in the circles: $0.384 + 0.095 + 0.055 = 0.534$ (off by 0.001 due to a rounding error).

Income Range	$0-25k	$25k-50k	$50k-100k	$100k+
(a)	0.18	0.39	0.33	0.16
(b)	0.38	-0.27	0.52	0.37
(c)	0.28	0.27	0.29	0.16

Figure 3.6: Proposed distributions of US household incomes (Guided Practice 3.16).

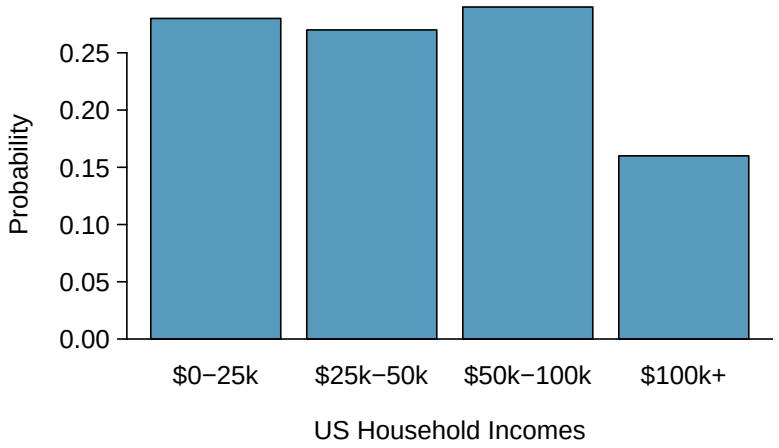


Figure 3.7: The probability distribution of US household income.

GUIDED PRACTICE 3.16

Figure 3.6 suggests three distributions for household income in the United States. Only one is correct. Which one must it be? What is wrong with the other two?[12]

Chapter 1 emphasized the importance of plotting data to provide quick summaries. Probability distributions can also be summarized in a bar plot. For instance, the distribution of US household incomes is shown in Figure 3.7 as a bar plot. The probability distribution for the sum of two dice is shown in Figure 3.5 and plotted in Figure 3.8.

In these bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a bar plot that resembles a

[12]The probabilities of (a) do not sum to 1. The second probability in (b) is negative. This leaves (c), which sure enough satisfies the requirements of a distribution. One of the three was said to be the actual distribution of US household incomes, so it must be (c).

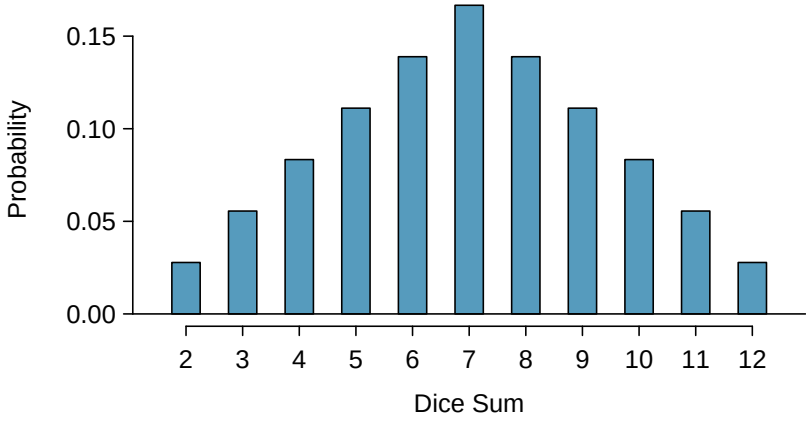


Figure 3.8: The probability distribution of the sum of two dice.

histogram, as in the case of the sum of two dice. Another example of plotting the bars at their respective locations is shown in Figure 3.20 on page 85.

3.1.6 Complement of an event

Rolling a die produces a value in the set {1, 2, 3, 4, 5, 6}. This set of all possible outcomes is called the **sample space** (S) for rolling a die. We often use the sample space to examine the scenario where an event does not occur.

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D. Figure 3.9 shows the relationship between D, D^c, and the sample space S.

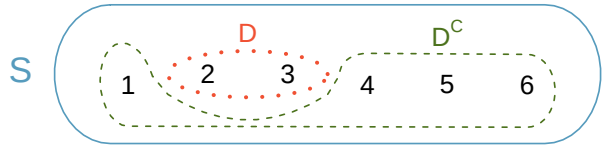


Figure 3.9: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible outcomes.

GUIDED PRACTICE 3.17

(a) Compute $P(D^c) = P(\text{rolling a } 1, 4, 5, \text{ or } 6)$. (b) What is $P(D) + P(D^c)$?[13]

GUIDED PRACTICE 3.18

Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 3.2 on page 61. (a) Write out what A^c and B^c represent. (b) Compute $P(A^c)$ and $P(B^c)$. (c) Compute $P(A) + P(A^c)$ and $P(B) + P(B^c)$.[14]

A complement of an event A is constructed to have two very important properties: (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. Property (i) implies

$$P(A \text{ or } A^c) = 1$$

That is, if the outcome is not in A, it must be represented in A^c. We use the Addition Rule for disjoint events to apply Property (ii):

$$P(A \text{ or } A^c) = P(A) + P(A^c)$$

Combining the last two equations yields a very useful relationship between the probability of an event and its complement.

COMPLEMENT

The complement of event A is denoted A^c, and A^c represents all outcomes not in A. A and A^c are mathematically related:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c)$$

[13] (a) The outcomes are disjoint and each has probability 1/6, so the total probability is 4/6 = 2/3. (b) We can also see that $P(D) = \frac{1}{6} + \frac{1}{6} = 1/3$. Since D and D^c are disjoint, $P(D) + P(D^c) = 1$.

[14] Brief solutions: (a) $A^c = \{3, 4, 5, 6\}$ and $B^c = \{1, 2, 3, 5\}$. (b) Noting that each outcome is disjoint, add the individual outcome probabilities to get $P(A^c) = 2/3$ and $P(B^c) = 2/3$. (c) A and A^c are disjoint, and the same is true of B and B^c. Therefore, $P(A) + P(A^c) = 1$ and $P(B) + P(B^c) = 1$.

In simple examples, computing A or A^c is feasible in a few steps. However, using the complement can save a lot of time as problems grow in complexity.

GUIDED PRACTICE 3.19

Let A represent the event where we roll two dice and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Figure 3.5 on page 63. (c) Determine $P(A)$.[15]

GUIDED PRACTICE 3.20

Find the following probabilities for rolling two dice:[16]

(a) The sum of the dice is *not* 6.

(b) The sum is at least 4. That is, determine the probability of the event $B = \{4, 5, ..., 12\}$.

(c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, ..., 10\}$.

3.1.7 Independence

Just as variables and observations can be independent, random processes can be independent, too. Two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

Example 3.5 provides a basic example of two independent processes: rolling two dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die. We first encountered this same question in Example 3.5 (page 58), where we calculated the probability using the following reasoning: 1/6 of the time the red die is a 1, and 1/6 of *those* times the white die will also be 1. This is illustrated in Figure 3.10. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6) \times (1/6) = 1/36$. This can be generalized to many independent processes.

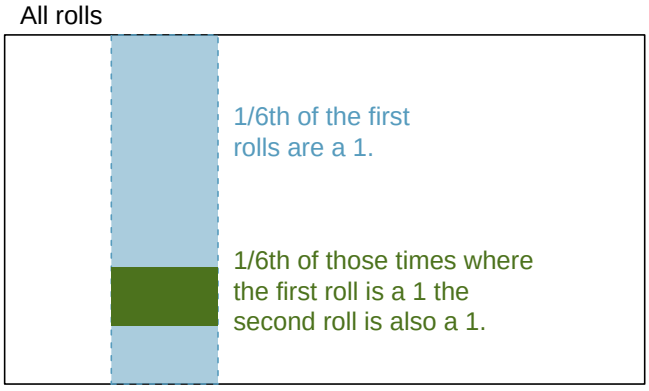


Figure 3.10: 1/6 of the time, the first roll is a 1. Then 1/6 of *those* times, the second roll will also be a 1.

[15] (a) The complement of A: when the total is equal to 12. (b) $P(A^c) = 1/36$. (c) Use the probability of the complement from part (b), $P(A^c) = 1/36$, and the equation for the complement: $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

[16] (a) First find $P(6) = 5/36$, then use the complement: $P(\text{not } 6) = 1 - P(6) = 31/36$.

(b) First find the complement, which requires much less effort: $P(2 \text{ or } 3) = 1/36 + 2/36 = 1/12$. Then calculate $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$.

(c) As before, finding the complement is the clever way to determine $P(D)$. First find $P(D^c) = P(11 \text{ or } 12) = 2/36 + 1/36 = 1/12$. Then calculate $P(D) = 1 - P(D^c) = 11/12$.

EXAMPLE 3.21

What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all 1s?

E

The same logic applies from Example 3.5. If 1/36 of the time the white and red dice are both 1, then 1/6 of *those* times the blue die will also be 1, so multiply:

$$\begin{aligned} P(white = 1 \text{ and } red = 1 \text{ and } blue = 1) &= P(white = 1) \times P(red = 1) \times P(blue = 1) \\ &= (1/6) \times (1/6) \times (1/6) = 1/216 \end{aligned}$$

Example 3.21 illustrates what is called the Multiplication Rule for independent processes.

MULTIPLICATION RULE FOR INDEPENDENT PROCESSES

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

GUIDED PRACTICE 3.22

G

About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?[17]

GUIDED PRACTICE 3.23

Suppose 5 people are selected at random.[18]

G

(a) What is the probability that all are right-handed?

(b) What is the probability that all are left-handed?

(c) What is the probability that not all of the people are right-handed?

Suppose the variables `handedness` and `sex` are independent, i.e. knowing someone's `sex` provides no useful information about their `handedness` and vice-versa. Then we can compute whether

[17] (a) The probability the first person is left-handed is 0.09, which is the same for the second person. We apply the Multiplication Rule for independent processes to determine the probability that both will be left-handed: $0.09 \times 0.09 = 0.0081$.

(b) It is reasonable to assume the proportion of people who are ambidextrous (both right and left handed) is nearly 0, which results in $P(\text{right-handed}) = 1 - 0.09 = 0.91$. Using the same reasoning as in part (a), the probability that both will be right-handed is $0.91 \times 0.91 = 0.8281$.

[18] (a) The abbreviations `RH` and `LH` are used for right-handed and left-handed, respectively. Since each are independent, we apply the Multiplication Rule for independent processes:

$$\begin{aligned} P(\text{all five are RH}) &= P(\text{first = RH, second = RH, ..., fifth = RH}) \\ &= P(\text{first = RH}) \times P(\text{second = RH}) \times \cdots \times P(\text{fifth = RH}) \\ &= 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624 \end{aligned}$$

(b) Using the same reasoning as in (a), $0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$

(c) Use the complement, $P(\text{all five are RH})$, to answer this question:

$$P(\text{not all RH}) = 1 - P(\text{all RH}) = 1 - 0.624 = 0.376$$

a randomly selected person is right-handed and female[19] using the Multiplication Rule:

$$\begin{aligned} P(\text{right-handed and female}) &= P(\text{right-handed}) \times P(\text{female}) \\ &= 0.91 \times 0.50 = 0.455 \end{aligned}$$

GUIDED PRACTICE 3.24

Three people are selected at random.[20]

(a) What is the probability that the first person is male and right-handed?

(b) What is the probability that the first two people are male and right-handed?.

(c) What is the probability that the third person is female and left-handed?

(d) What is the probability that the first two people are male and right-handed and the third person is female and left-handed?

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy $P(A \text{ and } B) = P(A) \times P(B)$.

EXAMPLE 3.25

If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is $1/4$ and the probability that it is an ace is $1/13$. The probability the card is the ace of hearts is $1/52$. We check whether $P(A \text{ and } B) = P(A) \times P(B)$ is satisfied:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

Because the equation holds, the event that the card is a heart and the event that the card is an ace are independent events.

[19] The actual proportion of the U.S. population that is `female` is about 50%, and so we use 0.5 for the probability of sampling a woman. However, this probability does differ in other countries.

[20] Brief answers are provided. (a) This can be written in probability notation as P(a randomly selected person is male and right-handed) = 0.455. (b) 0.207. (c) 0.045. (d) 0.0093.

3.2 Conditional probability

[Need to add an introductory paragraph.]

3.2.1 Exploring probabilities with a contingency table

The `photo_classify` data set represents a classifier a sample of 1822 photos from photo sharing website. Data scientists have been working to improve a classifier for whether the photo is about fashion or not, and these 659 photos represent a test for their classifier. Each photo gets two classifications: the first is called `mach_learn` and gives a classification from a machine learning (ML) system of either `pred_fashion` or `pred_not`. Each of these 1822 photos have also been classified carefully by a team of people, which we take to be the source of truth; this variable is called `truth` and takes values `fashion` and `not`. Figure 3.11 summarizes the results.

		truth		
		`fashion`	`not`	Total
`pred_not`	`pred_fashion`	197	22	219
	`not`	112	1491	1603
	Total	309	1513	1822

Figure 3.11: Contingency table summarizing the `photo_classify` data set.

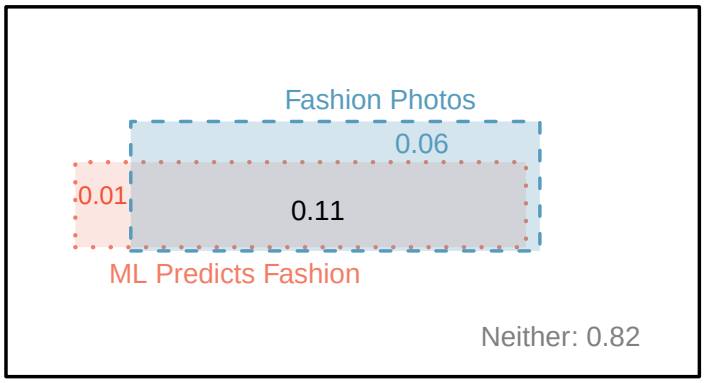


Figure 3.12: A Venn diagram using boxes for the `photo_classify` data set.

EXAMPLE 3.26

If a photo is actually about fashion, what is the chance the ML classifier correctly identified the photo as being about fashion?

E

We can estimate this probability using the data. Of the 309 fashion photos, the ML algorithm correctly classified 197 of the photos:

$$P(\texttt{mach\_learn} \text{ is } \texttt{pred\_fashion} \text{ given } \texttt{truth} \text{ is } \texttt{fashion}) = \frac{197}{309} = 0.638$$

EXAMPLE 3.27

We sample a photo from the data set and learn the ML algorithm predicted this photo was not about fashion. What is the probability that it was incorrect and the photo is about fashion?

(E) If the ML classifier suggests a photo is not about fashion, then it comes from the second row in the data set. Of these 1603 photos, 112 were actually about fashion:

$$P(\texttt{truth is fashion given mach\_learn is pred\_not}) = \frac{112}{1603} = 0.070$$

3.2.2 Marginal and joint probabilities

Figure 3.11 includes row and column totals for each variable separately in the `photo_classify` data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without regard to any other variables. For instance, a probability based solely on the `mach_learn` variable is a marginal probability:

$$P(\texttt{mach\_learn is pred\_fashion}) = \frac{219}{1822} = 0.12$$

A probability of outcomes for two or more variables or processes is called a **joint probability**:

$$P(\texttt{mach\_learn is pred\_fashion and truth is fashion}) = \frac{197}{1822} = 0.11$$

It is common to substitute a comma for "and" in a joint probability, although using either the word "and" or a comma is acceptable. That is,

$$P(\texttt{mach\_learn is pred\_fashion, truth is fashion})$$

means the same thing as

$$P(\texttt{mach\_learn is pred\_fashion and truth is fashion})$$

MARGINAL AND JOINT PROBABILITIES

If a probability is based on a single variable, it is a *marginal probability*. The probability of outcomes for two or more variables or processes is called a *joint probability*.

We use **table proportions** to summarize joint probabilities for the `photo_classify` sample. These proportions are computed by dividing each count in Figure 3.11 by the table's total, 1822, to obtain the proportions in Figure 3.13. The joint probability distribution of the `mach_learn` and `truth` variables is shown in Figure 3.14.

	truth: fashion	truth: not	Total
mach_learn: pred_fashion	0.1081	0.0121	0.1202
mach_learn: pred_not	0.0615	0.8183	0.8798
Total	0.1696	0.8304	1.00

Figure 3.13: Probability table summarizing the `photo_classify` data set.

GUIDED PRACTICE 3.28

(G) Verify Figure 3.14 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[21]

[21] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the

Joint outcome	Probability
`mach_learn` is `pred_fashion` and `truth` is `fashion`	0.1081
`mach_learn` is `pred_fashion` and `truth` is `not`	0.0121
`mach_learn` is `pred_not` and `truth` is `fashion`	0.0615
`mach_learn` is `pred_not` and `truth` is `not`	0.8183
Total	1.0000

Figure 3.14: Joint probability distribution for the `photo_classify` data set.

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability a randomly selected photo from the data set is about fashion is found by summing the outcomes where `truth` takes value `fashion`:

$$\begin{aligned} P(\underline{\texttt{truth is fashion}}) &= P(\texttt{mach\_learn is pred\_fashion and } \underline{\texttt{truth is fashion}}) \\ &\quad + P(\texttt{mach\_learn is pred\_not and } \underline{\texttt{truth is fashion}}) \\ &= 0.1081 + 0.0615 \\ &= 0.1696 \end{aligned}$$

3.2.3 Defining conditional probability

The ML classifier predicts whether a photo is about fashion, even if it is not perfect. We would like to better understand how to use information from a variable like `mach_learn` to improve our probability estimation of a second variable, which in this example is `truth`.

The probability that a random photo from the data set is about fashion is about 0.17. If we knew the machine learning classifier predicted the photo was about fashion, could we get a better estimate of the probability the photo is actually about fashion? Absolutely. To do so, we limit our view to only those 219 cases where the ML classifier predicted that the photo was about fashion and look at the fraction where the photo was actually about fashion:

$$P(\texttt{truth is fashion given mach\_learn is pred\_fashion}) = \frac{197}{219} = 0.900$$

We call this a **conditional probability** because we computed the probability under a condition: the ML classifier prediction said the photo was about fashion.

There are two parts to a conditional probability, the **outcome of interest** and the **condition**. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event. We generally separate the text inside our probability notation into the outcome of interest and the condition with a vertical bar:

$$\begin{aligned} &P(\texttt{truth is fashion given mach\_learn is pred\_fashion}) \\ &= P(\texttt{truth is fashion} \mid \texttt{mach\_learn is pred\_fashion}) = \frac{197}{219} = 0.900 \end{aligned}$$

The vertical bar "|" is read as *given*.

In the last equation, we computed the probability a photo was about fashion based on the condition that the ML algorithm predicted it was about fashion as a fraction:

$$\begin{aligned} &P(\texttt{truth is fashion} \mid \texttt{mach\_learn is pred\_fashion}) \\ &\quad = \frac{\#\text{ cases where } \underline{\texttt{truth is fashion}} \text{ and } \underline{\texttt{mach\_learn is pred\_fashion}}}{\#\text{ cases where } \texttt{truth is fashion}} \\ &\quad = \frac{197}{219} = 0.900 \end{aligned}$$

We considered only those cases that met the condition, `mach_learn` is `pred_fashion`, and then we computed the ratio of those cases that satisfied our outcome of interest, photo was actually about

probabilities is $0.1081 + 0.0121 + 0.0615 + 0.8183 = 1.00$.

fashion.

Frequently, marginal and joint probabilities are provided instead of count data. For example, disease rates are commonly listed in percentages rather than in a count format. We would like to be able to compute conditional probabilities even when no counts are available, and we use the last equation as a template to understand this technique.

We considered only those cases that satisfied the condition, where the ML algorithm predicted fashion. Of these cases, the conditional probability was the fraction representing the outcome of interest, that the photo was about fashion. Suppose we were provided only the information in Figure 3.13, i.e. only probability data. Then if we took a sample of 1000 photos, we would anticipate about 12.0% or $0.120 \times 1000 = 120$ would be predicted to be about fashion (`mach_learn` is `pred_fashion`). Similarly, we would expect about 10.8% or $0.108 \times 1000 = 108$ to meet both the information criteria and represent our outcome of interest. Then the conditional probability can be computed as

$$\begin{aligned} &P(\texttt{truth is fashion} \mid \texttt{mach\_learn is pred\_fashion}) \\ &= \frac{\#\ (\texttt{truth is fashion and mach\_learn is pred\_fashion})}{\#\ (\texttt{mach\_learn is pred\_fashion})} \\ &= \frac{108}{120} = \frac{0.108}{0.120} = 0.90 \end{aligned}$$

Here we are examining exactly the fraction of two probabilities, 0.108 and 0.120, which we can write as

$$P(\texttt{truth is fashion and mach\_learn is pred\_fashion}) \quad \text{and} \quad P(\texttt{mach\_learn is pred\_fashion}).$$

The fraction of these probabilities is an example of the general formula for conditional probability.

CONDITIONAL PROBABILITY

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

GUIDED PRACTICE 3.29

(a) Write out the following statement in conditional probability notation: "*The probability that the ML prediction was correct, if the photo was about fashion*". Here the condition is now based on the photo's `truth` status, not the ML algorithm.

(b) Determine the probability from part (a). Table 3.13 on page 70 may be helpful.[22]

[22] (a) If the photo is about fashion and the ML algorithm prediction was correct, then the ML algorithm my have a value of `pred_fashion`:

$$P(\texttt{mach\_learn is pred\_fashion} \mid \texttt{truth is fashion})$$

(b) The equation for conditional probability indicates we should first find $P(\texttt{mach\_learn is pred\_fashion and truth is fashion}) = 0.1081$ and $P(\texttt{truth is not}) = 0.1696$. Then the ratio represents the conditional probability: $0.1081/0.1696 = 0.6374$.

GUIDED PRACTICE 3.30

(a) Determine the probability that the algorithm is incorrect if it is known the photo is about fashion.

(b) Using the answers from part (a) and Guided Practice 3.29(b), compute

$$P(\texttt{mach\_learn is pred\_fashion} \mid \texttt{truth is fashion}) + P(\texttt{mach\_learn is pred\_not} \mid \texttt{truth is fashion})$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[23]

3.2.4 Smallpox in Boston, 1721

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston. Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: `inoculated` and `result`. The variable `inoculated` takes two levels: `yes` or `no`, indicating whether the person was inoculated or not. The variable `result` has outcomes `lived` or `died`. These data are summarized in Tables 3.15 and 3.16.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Figure 3.15: Contingency table for the `smallpox` data set.

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Figure 3.16: Table proportions for the `smallpox` data, computed by dividing each count by the table total, 6224.

GUIDED PRACTICE 3.31

Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[24]

GUIDED PRACTICE 3.32

Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 3.31?[25]

[23] (a) This probability is $\frac{P(\texttt{mach\_learn is pred\_not, truth is fashion})}{P(\texttt{truth is fashion})} = \frac{0.0615}{0.1696} = 0.3626$. (b) The total equals 1. (c) Under the condition the photo is about fashion, the ML algorithm must have either predicted it was about fashion or predicted it was not about fashion. The complement still works for conditional probabilities, provided the probabilities are conditioned on the same information.

[24] $P(\texttt{result = died} \mid \texttt{inoculated = no}) = \frac{P(\texttt{result = died and inoculated = no})}{P(\texttt{inoculated = no})} = \frac{0.1356}{0.9608} = 0.1411$.

[25] $P(\texttt{result = died} \mid \texttt{inoculated = yes}) = \frac{P(\texttt{result = died and inoculated = yes})}{P(\texttt{inoculated = yes})} = \frac{0.0010}{0.0392} = 0.0255$ (if we avoided rounding errors, we'd get $6/244 = 0.0246$). The death rate for individuals who were inoculated is only about 1 in 40

GUIDED PRACTICE 3.33

The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone `lived` or `died` and also affect whether that person was inoculated?[26]

3.2.5 General multiplication rule

Section 3.1.7 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

GENERAL MULTIPLICATION RULE

If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

It is useful to think of A as the outcome of interest and B as the condition.

This General Multiplication Rule is simply a rearrangement of the equation for conditional probability.

EXAMPLE 3.34

Consider the `smallpox` data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

We will compute our answer using the General Multiplication Rule and then verify it using Figure 3.16. We want to determine

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no})$$

and we are given that

$$P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{no}) = 0.8588$$
$$P(\texttt{inoculated} = \texttt{no}) = 0.9608$$

Among the 96.08% of people who were not inoculated, 85.88% survived:

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Figure 3.16 at the intersection of `no` and `lived` (with a small rounding error).

GUIDED PRACTICE 3.35

Use $P(\texttt{inoculated} = \texttt{yes}) = 0.0392$ and $P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{yes}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[27]

while the death rate is about 1 in 7 for those who were not inoculated.

[26] Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care. There are other valid answers for part (c).

[27] The answer is 0.0382, which can be verified using Figure 3.16.

GUIDED PRACTICE 3.36

If 97.54% of the people who were inoculated lived, what proportion of inoculated people must have died?[28]

SUM OF CONDITIONAL PROBABILITIES

Let A_1, ..., A_k represent all the disjoint outcomes for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$P(A_1|B) + \cdots + P(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$P(A|B) = 1 - P(A^c|B)$$

GUIDED PRACTICE 3.37

Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[29]

3.2.6 Independence considerations in conditional probability

If two events are independent, then knowing the outcome of one should provide no information about the other. We can show this is mathematically true using conditional probabilities.

GUIDED PRACTICE 3.38

Let X and Y represent the outcomes of rolling two dice.[30]

(a) What is the probability that the first die, X, is 1?

(b) What is the probability that both X and Y are 1?

(c) Use the formula for conditional probability to compute $P(Y = 1 \mid X = 1)$.

(d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.

We can show in Guided Practice 3.38(c) that the conditioning information has no influence by using the Multiplication Rule for independence processes:

$$\begin{aligned} P(Y = 1 \mid X = 1) &= \frac{P(Y = 1 \text{ and } X = 1)}{P(X = 1)} \\ &= \frac{P(Y = 1) \times P(X = 1)}{P(X = 1)} \\ &= P(Y = 1) \end{aligned}$$

[28] There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.54% = 2.46% of the people who were inoculated died.

[29] The samples are large relative to the difference in death rates for the "inoculated" and "not inoculated" groups, so it seems there is an association between `inoculated` and `outcome`. However, as noted in the solution to Guided Practice 3.33, this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

[30] Brief solutions: (a) 1/6. (b) 1/36. (c) $\frac{P(Y=1 \text{ and } X=1)}{P(X=1)} = \frac{1/36}{1/6} = 1/6$. (d) The probability is the same as in part (c): $P(Y = 1) = 1/6$. The probability that $Y = 1$ was unchanged by knowledge about X, which makes sense as X and Y are independent.

GUIDED PRACTICE 3.39

Ron is watching a roulette table in a casino and notices that the last five outcomes were `black`. He figures that the chances of getting `black` six times in a row is very small (about 1/64) and puts his paycheck on red. What is wrong with his reasoning?[31]

3.2.7 Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

The `smallpox` data fit this description. We see the population as split by `inoculation`: `yes` and `no`. Following this split, survival rates were observed for each group. This structure is reflected in the **tree diagram** shown in Figure 3.17. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

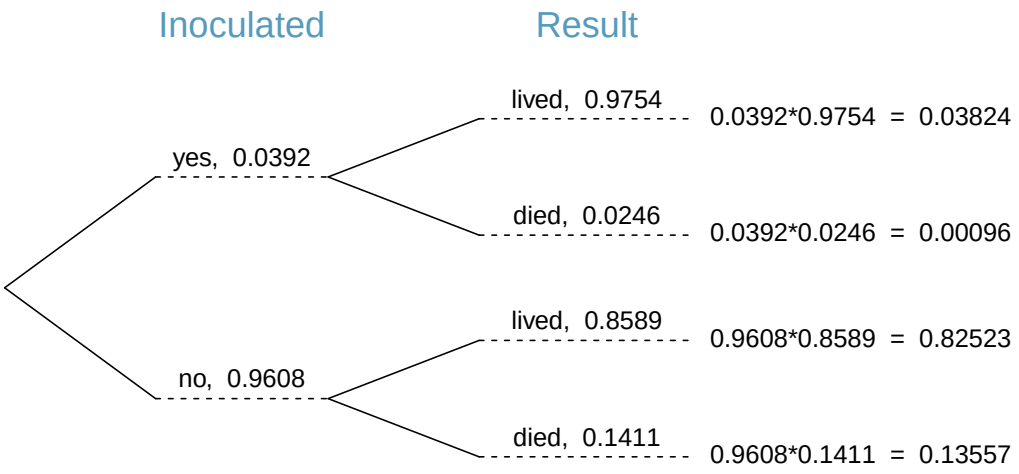


Figure 3.17: A tree diagram of the `smallpox` data set.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 3.17. This tree diagram splits the smallpox data by `inoculation` into the `yes` and `no` groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 3.17 is the probability that `result` = `lived` conditioned on the information that `inoculated` = `yes`. We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned} &P(\texttt{inoculated} = \texttt{yes} \text{ and } \texttt{result} = \texttt{lived}) \\ &\quad = P(\texttt{inoculated} = \texttt{yes}) \times P(\texttt{result} = \texttt{lived} | \texttt{inoculated} = \texttt{yes}) \\ &\quad = 0.0392 \times 0.9754 = 0.0382 \end{aligned}$$

[31] He has forgotten that the next roulette spin is independent of the previous spins. Casinos do employ this practice; they post the last several outcomes of many betting games to trick unsuspecting gamblers into believing the odds are in their favor. This is called the **gambler's fallacy**.

EXAMPLE 3.40

Consider the midterm and final for a statistics class. Suppose 13% of students earned an `A` on the midterm. Of those students who earned an `A` on the midterm, 47% received an `A` on the final, and 11% of the students who earned lower than an `A` on the midterm received an `A` on the final. You randomly pick up a final exam and notice the student received an `A`. What is the probability that this student earned an `A` on the midterm?

The end-goal is to find $P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A})$. To calculate this conditional probability, we need the following probabilities:

$$P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) \qquad \text{and} \qquad P(\texttt{final} = \texttt{A})$$

However, this information is not provided, and it is not obvious how to calculate these probabilities. Since we aren't sure how to proceed, it is useful to organize the information into a tree diagram, as shown in Figure 3.18. When constructing a tree diagram, variables provided with marginal probabilities are often used to create the tree's primary branches; in this case, the marginal probabilities are provided for midterm grades. The final grades, which correspond to the conditional probabilities provided, will be shown on the secondary branches.

E

With the tree diagram constructed, we may compute the required probabilities:

$$\begin{aligned} &P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) = 0.0611 \\ &P(\underline{\texttt{final} = \texttt{A}}) \\ &\quad = P(\texttt{midterm} = \texttt{other} \text{ and } \underline{\texttt{final} = \texttt{A}}) + P(\texttt{midterm} = \texttt{A} \text{ and } \underline{\texttt{final} = \texttt{A}}) \\ &\quad = 0.0957 + 0.0611 = 0.1568 \end{aligned}$$

The marginal probability, $P(\texttt{final} = \texttt{A})$, was calculated by adding up all the joint probabilities on the right side of the tree that correspond to $\texttt{final} = \texttt{A}$. We may now finally take the ratio of the two probabilities:

$$\begin{aligned} P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A}) &= \frac{P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A})}{P(\texttt{final} = \texttt{A})} \\ &= \frac{0.0611}{0.1568} = 0.3897 \end{aligned}$$

The probability the student also earned an A on the midterm is about 0.39.

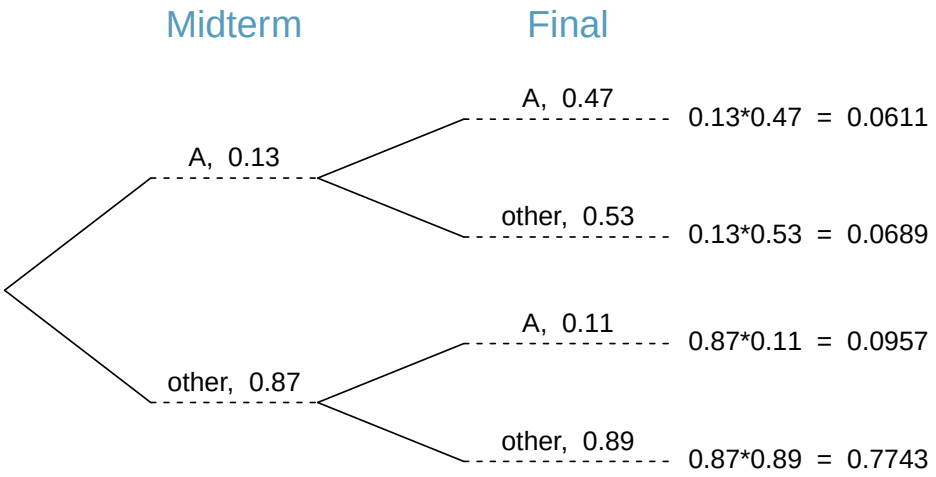


Figure 3.18: A tree diagram describing the `midterm` and `final` variables.

GUIDED PRACTICE 3.41

After an introductory statistics course, 78% of students can successfully construct tree diagrams. Of those who can construct tree diagrams, 97% passed, while only 57% of those students who could not construct tree diagrams passed. (a) Organize this information into a tree diagram. (b) What is the probability that a randomly selected student passed? (c) Compute the probability a student is able to construct a tree diagram if it is known that she passed.[32]

3.2.8 Bayes' Theorem

In many instances, we are given a conditional probability of the form

$$P(\text{statement about variable 1} \mid \text{statement about variable 2})$$

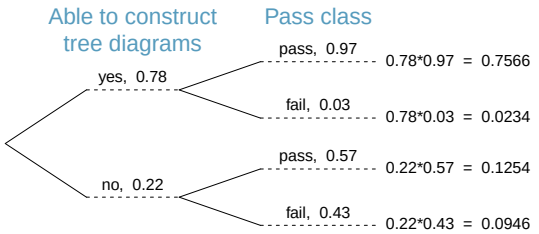
but we would really like to know the inverted conditional probability:

$$P(\text{statement about variable 2} \mid \text{statement about variable 1})$$

Tree diagrams can be used to find the second conditional probability when given the first. However, sometimes it is not possible to draw the scenario in a tree diagram. In these cases, we can apply a very useful and general formula: Bayes' Theorem.

We first take a critical look at an example of inverting conditional probabilities where we still apply a tree diagram.

[32] (a) The tree diagram is shown to the right. (b) Identify which two joint probabilities represent students who passed, and add them: $P(\text{passed}) = 0.7566 + 0.1254 = 0.8820$. (c) $P(\text{construct tree diagram} \mid \text{passed}) = \frac{0.7566}{0.8820} = 0.8578$.



EXAMPLE 3.42

In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but this test is not perfect. In about 11% of patients with breast cancer, the test gives a **false negative**: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a **false positive** in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not. If we tested a random woman over 40 for breast cancer using a mammogram and the test came back positive – that is, the test suggested the patient has cancer – what is the probability that the patient actually has breast cancer?

Notice that we are given sufficient information to quickly compute the probability of testing positive if a woman has breast cancer ($1.00 - 0.11 = 0.89$). However, we seek the inverted probability of cancer given a positive test result. (Watch out for the non-intuitive medical language: a *positive* test result suggests the possible presence of cancer in a mammogram screening.) This inverted probability may be broken into two pieces:

$$P(\text{has BC} \mid \text{mammogram}^+) = \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)}$$

where "has BC" is an abbreviation for the patient actually having breast cancer and "mammogram$^+$" means the mammogram screening was positive. A tree diagram is useful for identifying each probability and is shown in Figure 3.19. The probability the patient has breast cancer and the mammogram is positive is

(E)

$$\begin{aligned} P(\text{has BC and mammogram}^+) &= P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC}) \\ &= 0.89 \times 0.0035 = 0.00312 \end{aligned}$$

The probability of a positive test result is the sum of the two corresponding scenarios:

$$\begin{aligned} P(\underline{\text{mammogram}^+}) &= P(\underline{\text{mammogram}^+} \text{ and has BC}) \\ &\quad + P(\underline{\text{mammogram}^+} \text{ and no BC}) \\ &= P(\text{has BC})P(\text{mammogram}^+ \mid \text{has BC}) \\ &\quad + P(\text{no BC})P(\text{mammogram}^+ \mid \text{no BC}) \\ &= 0.0035 \times 0.89 + 0.9965 \times 0.07 = 0.07288 \end{aligned}$$

Then if the mammogram screening is positive for a patient, the probability the patient has breast cancer is

$$\begin{aligned} P(\text{has BC} \mid \text{mammogram}^+) &= \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)} \\ &= \frac{0.00312}{0.07288} \approx 0.0428 \end{aligned}$$

That is, even if a patient has a positive mammogram screening, there is still only a 4% chance that she has breast cancer.

Example 3.42 highlights why doctors often run more tests regardless of a first positive test result. When a medical condition is rare, a single positive test isn't generally definitive.

Consider again the last equation of Example 3.42. Using the tree diagram, we can see that the numerator (the top of the fraction) is equal to the following product:

$$P(\text{has BC and mammogram}^+) = P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})$$

The denominator – the probability the screening was positive – is equal to the sum of probabilities for each positive screening scenario:

$$P(\underline{\text{mammogram}^+}) = P(\underline{\text{mammogram}^+} \text{ and no BC}) + P(\underline{\text{mammogram}^+} \text{ and has BC})$$

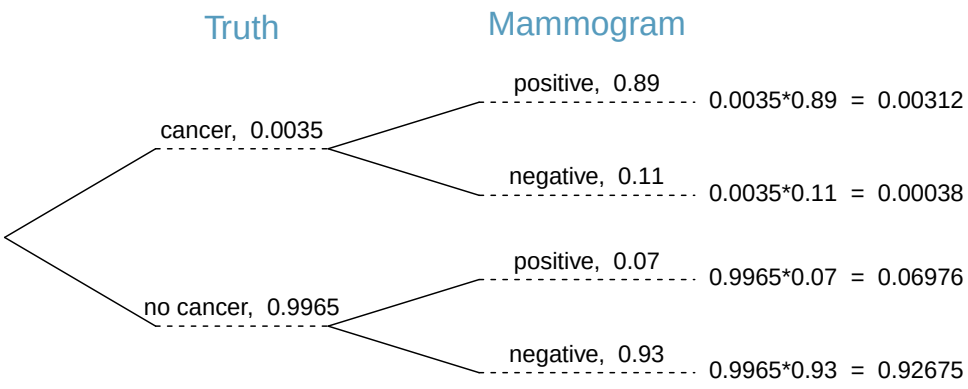


Figure 3.19: Tree diagram for Example 3.42, computing the probability a random patient who tests positive on a mammogram actually has breast cancer.

In the example, each of the probabilities on the right side was broken down into a product of a conditional probability and marginal probability using the tree diagram.

$$\begin{aligned} P(\text{mammogram}^+) &= P(\text{mammogram}^+ \text{ and no BC}) + P(\text{mammogram}^+ \text{ and has BC}) \\ &= P(\text{mammogram}^+ \mid \text{no BC})P(\text{no BC}) \\ &\quad + P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC}) \end{aligned}$$

We can see an application of Bayes' Theorem by substituting the resulting probability expressions into the numerator and denominator of the original conditional probability.

$$\begin{aligned} &P(\text{has BC} \mid \text{mammogram}^+) \\ &\quad = \frac{P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})}{P(\text{mammogram}^+ \mid \text{no BC})P(\text{no BC}) + P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})} \end{aligned}$$

BAYES' THEOREM: INVERTING PROBABILITIES

Consider the following conditional probability for variable 1 and variable 2:

$$P(\text{outcome } A_1 \text{ of variable 1} \mid \text{outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable.

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The numerator identifies the probability of getting both A_1 and B. The denominator is the marginal probability of getting B. This bottom component of the fraction appears long and complicated since we have to add up probabilities from all of the different ways to get B. We always completed this step when using tree diagrams. However, we usually did it in a separate step so it didn't seem as complex.

To apply Bayes' Theorem correctly, there are two preparatory steps:

(1) First identify the marginal probabilities of each possible outcome of the first variable: $P(A_1)$, $P(A_2)$, ..., $P(A_k)$.

(2) Then identify the probability of the outcome B, conditioned on each possible scenario for the first variable: $P(B|A_1)$, $P(B|A_2)$, ..., $P(B|A_k)$.

Once each of these probabilities are identified, they can be applied directly within the formula.

TIP: Only use Bayes' Theorem when tree diagrams are difficult
Drawing a tree diagram makes it easier to understand how two variables are connected. Use Bayes' Theorem only when there are so many scenarios that drawing a tree diagram would be complex.

GUIDED PRACTICE 3.43

Jose visits campus every Thursday evening. However, some days the parking garage is full, often due to college events. There are academic events on 35% of evenings, sporting events on 20% of evenings, and no events on 45% of evenings. When there is an academic event, the garage fills up about 25% of the time, and it fills up 70% of evenings with sporting events. On evenings when there are no events, it only fills up about 5% of the time. If Jose comes to campus and finds the garage full, what is the probability that there is a sporting event? Use a tree diagram to solve this problem.[33]

EXAMPLE 3.44

Here we solve the same problem presented in Guided Practice 3.43, except this time we use Bayes' Theorem.

The outcome of interest is whether there is a sporting event (call this A_1), and the condition is that the lot is full (B). Let A_2 represent an academic event and A_3 represent there being no event on campus. Then the given probabilities can be written as

$$P(A_1) = 0.2 \qquad P(A_2) = 0.35 \qquad P(A_3) = 0.45$$
$$P(B|A_1) = 0.7 \qquad P(B|A_2) = 0.25 \qquad P(B|A_3) = 0.05$$

Bayes' Theorem can be used to compute the probability of a sporting event (A_1) under the condition that the parking lot is full (B):

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.7)(0.2)}{(0.7)(0.2) + (0.25)(0.35) + (0.05)(0.45)} \\ &= 0.56 \end{aligned}$$

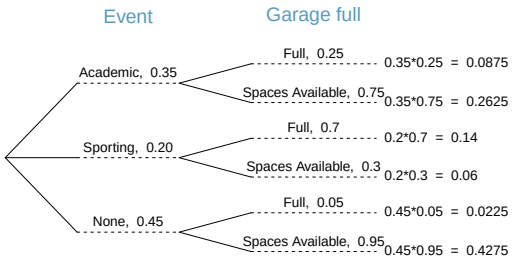
Based on the information that the garage is full, there is a 56% probability that a sporting event is being held on campus that evening.

GUIDED PRACTICE 3.45

Use the information in the previous exercise and example to verify the probability that there is an academic event conditioned on the parking lot being full is 0.35.[34]

---

[33] The tree diagram, with three primary branches, is shown to the right. Next, we identify two probabilities from the tree diagram. (1) The probability that there is a sporting event and the garage is full: 0.14. (2) The probability the garage is full: $0.0875 + 0.14 + 0.0225 = 0.25$. Then the solution is the ratio of these probabilities: $\frac{0.14}{0.25} = 0.56$. If the garage is full, there is a 56% probability that there is a sporting event.



GUIDED PRACTICE 3.46

G In Guided Practice 3.43 and 3.45, you found that if the parking lot is full, the probability there is a sporting event is 0.56 and the probability there is an academic event is 0.35. Using this information, compute $P(\text{no event} \mid \text{the lot is full})$.[35]

The last several exercises offered a way to update our belief about whether there is a sporting event, academic event, or no event going on at the school based on the information that the parking lot was full. This strategy of *updating beliefs* using Bayes' Theorem is actually the foundation of an entire section of statistics called **Bayesian statistics**. While Bayesian statistics is very important and useful, we will not have time to cover much more of it in this book.

[34]Short answer:

$$\begin{aligned} P(A_2|B) &= \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.25)(0.35)}{(0.7)(0.2) + (0.25)(0.35) + (0.05)(0.45)} \\ &= 0.35 \end{aligned}$$

[35]Each probability is conditioned on the same information that the garage is full, so the complement may be used: $1.00 - 0.56 - 0.35 = 0.09$.

3.3 Sampling from a small population

EXAMPLE 3.47

Professors sometimes select a student at random to answer a question. If each student has an equal chance of being selected and there are 15 people in your class, what is the chance that she will pick you for the next question?

(E)

If there are 15 people to ask and none are skipping class, then the probability is 1/15, or about 0.067.

EXAMPLE 3.48

If the professor asks 3 questions, what is the probability that you will not be selected? Assume that she will not pick the same person twice in a given lecture.

For the first question, she will pick someone else with probability 14/15. When she asks the second question, she only has 14 people who have not yet been asked. Thus, if you were not picked on the first question, the probability you are again not picked is 13/14. Similarly, the probability you are again not picked on the third question is 12/13, and the probability of not being picked for any of the three questions is

(E)

$$\begin{aligned} &P(\text{not picked in 3 questions}) \\ &\quad = P(\texttt{Q1} = \texttt{not\_picked}, \texttt{Q2} = \texttt{not\_picked}, \texttt{Q3} = \texttt{not\_picked}.) \\ &\quad = \frac{14}{15} \times \frac{13}{14} \times \frac{12}{13} = \frac{12}{15} = 0.80 \end{aligned}$$

GUIDED PRACTICE 3.49

(G)

What rule permitted us to multiply the probabilities in Example 3.48?[36]

EXAMPLE 3.50

Suppose the professor randomly picks without regard to who she already selected, i.e. students can be picked more than once. What is the probability that you will not be picked for any of the three questions?

Each pick is independent, and the probability of not being picked for any individual question is 14/15. Thus, we can use the Multiplication Rule for independent processes.

(E)

$$\begin{aligned} &P(\text{not picked in 3 questions}) \\ &\quad = P(\texttt{Q1} = \texttt{not\_picked}, \texttt{Q2} = \texttt{not\_picked}, \texttt{Q3} = \texttt{not\_picked}.) \\ &\quad = \frac{14}{15} \times \frac{14}{15} \times \frac{14}{15} = 0.813 \end{aligned}$$

You have a slightly higher chance of not being picked compared to when she picked a new person for each question. However, you now may be picked more than once.

[36] The three probabilities we computed were actually one marginal probability, $P(\texttt{Q1=not\_picked})$, and two conditional probabilities:

$$P(\texttt{Q2} = \texttt{not\_picked} \mid \texttt{Q1} = \texttt{not\_picked})$$
$$P(\texttt{Q3} = \texttt{not\_picked} \mid \texttt{Q1} = \texttt{not\_picked}, \texttt{Q2} = \texttt{not\_picked})$$

Using the General Multiplication Rule, the product of these three probabilities is the probability of not being picked in 3 questions.

GUIDED PRACTICE 3.51

Under the setup of Example 3.50, what is the probability of being picked to answer all three questions?[37]

If we sample from a small population **without replacement**, we no longer have independence between our observations. In Example 3.48, the probability of not being picked for the second question was conditioned on the event that you were not picked for the first question. In Example 3.50, the professor sampled her students **with replacement**: she repeatedly sampled the entire class without regard to who she already picked.

GUIDED PRACTICE 3.52

Your department is holding a raffle. They sell 30 tickets and offer seven prizes. (a) They place the tickets in a hat and draw one for each prize. The tickets are sampled without replacement, i.e. the selected tickets are not placed back in the hat. What is the probability of winning a prize if you buy one ticket? (b) What if the tickets are sampled with replacement?[38]

GUIDED PRACTICE 3.53

Compare your answers in Guided Practice 3.52. How much influence does the sampling method have on your chances of winning a prize?[39]

Had we repeated Guided Practice 3.52 with 300 tickets instead of 30, we would have found something interesting: the results would be nearly identical. The probability would be 0.0233 without replacement and 0.0231 with replacement. When the sample size is only a small fraction of the population (under 10%), observations are nearly independent even when sampling without replacement.

[37] $P(\text{being picked to answer all three questions}) = \left(\frac{1}{15}\right)^3 = 0.00030$.

[38] (a) First determine the probability of not winning. The tickets are sampled without replacement, which means the probability you do not win on the first draw is 29/30, 28/29 for the second, ..., and 23/24 for the seventh. The probability you win no prize is the product of these separate probabilities: 23/30. That is, the probability of winning a prize is $1 - 23/30 = 7/30 = 0.233$. (b) When the tickets are sampled with replacement, there are seven independent draws. Again we first find the probability of not winning a prize: $(29/30)^7 = 0.789$. Thus, the probability of winning (at least) one prize when drawing with replacement is 0.211.

[39] There is about a 10% larger chance of winning a prize when using sampling without replacement. However, at most one prize may be won under this sampling procedure.

3.4 Random variables

EXAMPLE 3.54

Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one term to another. If there are 100 students enrolled, how many books should the bookstore expect to sell to this class?

(E)

Around 20 students will not buy either book (0 books total), about 55 will buy one book (55 books total), and approximately 25 will buy two books (totaling 50 books for these 25 students). The bookstore should expect to sell about 105 books for this class.

(G) **GUIDED PRACTICE 3.55**

Would you be surprised if the bookstore sold slightly more or less than 105 books?[40]

EXAMPLE 3.56

The textbook costs \$137 and the study guide \$33. How much revenue should the bookstore expect from this class of 100 students?

(E)

About 55 students will just buy a textbook, providing revenue of

$$\$137 \times 55 = \$7,535$$

The roughly 25 students who buy both the textbook and the study guide would pay a total of

$$(\$137 + \$33) \times 25 = \$170 \times 25 = \$4,250$$

Thus, the bookstore should expect to generate about $\$7,535 + \$4,250 = \$11,785$ from these 100 students for this one class. However, there might be some *sampling variability* so the actual amount may differ by a little bit.

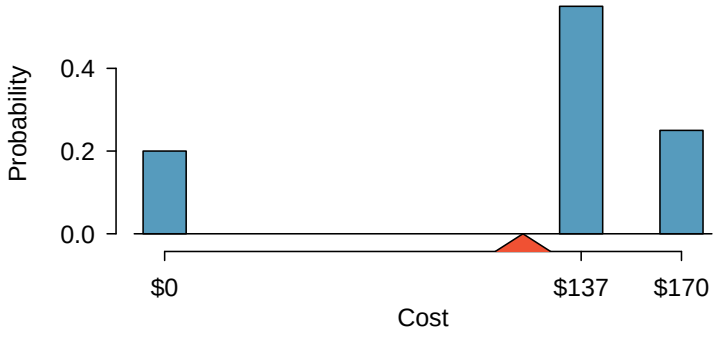


Figure 3.20: Probability distribution for the bookstore's revenue from a single student. The distribution balances on a triangle representing the average revenue per student.

[40] If they sell a little more or a little less, this should not be a surprise. Hopefully Chapter 1 helped make clear that there is natural variability in observed data. For example, if we would flip a coin 100 times, it will not usually come up heads exactly half the time, but it will probably be close.

EXAMPLE 3.57

What is the average revenue per student for this course?

(E)

The expected total revenue is \$11,785, and there are 100 students. Therefore the expected revenue per student is $\$11,785/100 = \$117.85$.

3.4.1 Expectation

We call a variable or process with a numerical outcome a **random variable**, and we usually represent this random variable with a capital letter such as X, Y, or Z. The amount of money a single student will spend on her statistics books is a random variable, and we represent it by X.

RANDOM VARIABLE

A random process or variable with a numerical outcome.

The possible outcomes of X are labeled with a corresponding lower case letter x and subscripts. For example, we write $x_1 = \$0$, $x_2 = \$137$, and $x_3 = \$170$, which occur with probabilities 0.20, 0.55, and 0.25. The distribution of X is summarized in Figure 3.20 and Figure 3.21.

i	1	2	3	Total
x_i	\$0	\$137	\$170	–
$P(X = x_i)$	0.20	0.55	0.25	1.00

Figure 3.21: The probability distribution for the random variable X, representing the bookstore's revenue from a single student.

We computed the average outcome of X as \$117.85 in Example 3.57. We call this average the **expected value** of X, denoted by $E(X)$. The expected value of a random variable is computed by adding each outcome weighted by its probability:

$$\begin{aligned} E(X) &= 0 \times P(X = 0) + 137 \times P(X = 137) + 170 \times P(X = 170) \\ &= 0 \times 0.20 + 137 \times 0.55 + 170 \times 0.25 = 117.85 \end{aligned}$$

EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + \cdots + x_k \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \end{aligned}$$

The Greek letter μ may be used in place of the notation $E(X)$.

The expected value for a random variable represents the average outcome. For example, $E(X) = 117.85$ represents the average amount the bookstore expects to make from a single student, which we could also write as $\mu = 117.85$.

It is also possible to compute the expected value of a continuous random variable (see Section 3.5). However, it requires a little calculus and we save it for a later class.[41]

In physics, the expectation holds the same meaning as the center of gravity. The distribution can be represented by a series of weights at each outcome, and the mean represents the balancing point. This is represented in Figures 3.20 and 3.22. The idea of a center of gravity also expands

[41] $\mu = \int x f(x) dx$ where $f(x)$ represents a function for the density curve.

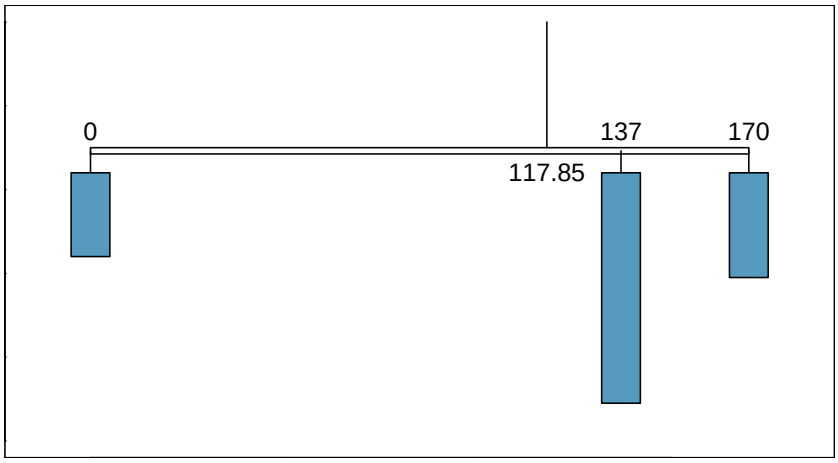


Figure 3.22: A weight system representing the probability distribution for X. The string holds the distribution at the mean to keep the system balanced.

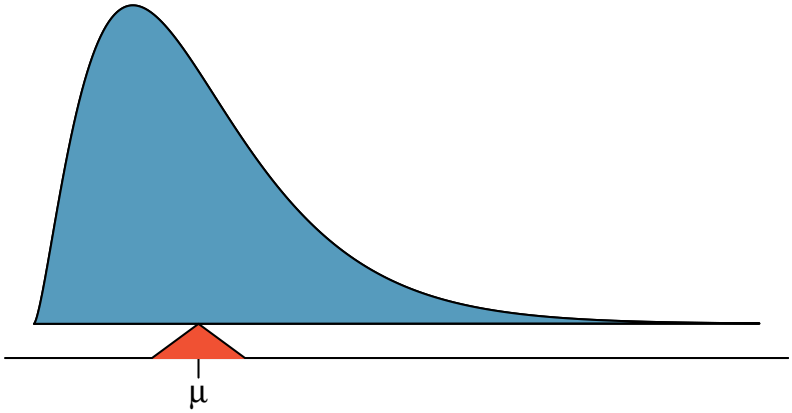


Figure 3.23: A continuous distribution can also be balanced at its mean.

to continuous probability distributions. Figure 3.23 shows a continuous probability distribution balanced atop a wedge placed at the mean.

3.4.2 Variability in random variables

Suppose you ran the university bookstore. Besides how much revenue you expect to generate, you might also want to know the volatility (variability) in your revenue.

The variance and standard deviation can be used to describe the variability of a random variable. Section 2.1.4 introduced a method for finding the variance and standard deviation for a data set. We first computed deviations from the mean ($x_i - \mu$), squared those deviations, and took an average to get the variance. In the case of a random variable, we again compute squared deviations. However, we take their sum weighted by their corresponding probabilities, just like we did for the expectation. This weighted sum of squared deviations equals the variance, and we calculate the standard deviation by taking the square root of the variance, just as we did in Section 2.1.4.

GENERAL VARIANCE FORMULA

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by $Var(X)$ or the symbol σ^2, is

$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 \times P(X = x_1) + \cdots \\ &\qquad \cdots + (x_k - \mu)^2 \times P(X = x_k) \\ &= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j)\end{aligned}$$

The standard deviation of X, labeled σ, is the square root of the variance.

EXAMPLE 3.58

Compute the expected value, variance, and standard deviation of X, the revenue of a single statistics student for the bookstore.

It is useful to construct a table that holds computations for each outcome separately, then add up the results.

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \times P(X = x_i)$	0	75.35	42.50	117.85

E

Thus, the expected value is $\mu = 117.85$, which we computed earlier. The variance can be constructed by extending this table:

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \times P(X = x_i)$	0	75.35	42.50	117.85
$x_i - \mu$	-117.85	19.15	52.15	
$(x_i - \mu)^2$	13888.62	366.72	2719.62	
$(x_i - \mu)^2 \times P(X = x_i)$	2777.7	201.7	679.9	3659.3

The variance of X is $\sigma^2 = 3659.3$, which means the standard deviation is $\sigma = \sqrt{3659.3} = \$60.49$.

GUIDED PRACTICE 3.59

The bookstore also offers a chemistry textbook for \$159 and a book supplement for \$41. From past experience, they know about 25% of chemistry students just buy the textbook while 60% buy both the textbook and supplement.[42]

(a) What proportion of students don't buy either book? Assume no students buy the supplement without the textbook.

G

(b) Let Y represent the revenue from a single student. Write out the probability distribution of Y, i.e. a table for each outcome and its associated probability.

(c) Compute the expected revenue from a single chemistry student.

(d) Find the standard deviation to describe the variability associated with the revenue from a single student.

[42] (a) 100% - 25% - 60% = 15% of students do not buy any books for the class. Part (b) is represented by the first two lines in the table below. The expectation for part (c) is given as the total on the line $y_i \times P(Y = y_i)$. The result of part (d) is the square-root of the variance listed on in the total on the last line: $\sigma = \sqrt{Var(Y)} = \$69.28$.

3.4.3 Linear combinations of random variables

So far, we have thought of each variable as being a complete story in and of itself. Sometimes it is more appropriate to use a combination of variables. For instance, the amount of time a person spends commuting to work each week can be broken down into several daily commutes. Similarly, the total gain or loss in a stock portfolio is the sum of the gains and losses in its components.

EXAMPLE 3.60

John travels to work five days a week. We will use X_1 to represent his travel time on Monday, X_2 to represent his travel time on Tuesday, and so on. Write an equation using X_1, ..., X_5 that represents his travel time for the week, denoted by W.

E His total weekly travel time is the sum of the five daily values:

$$W = X_1 + X_2 + X_3 + X_4 + X_5$$

Breaking the weekly travel time W into pieces provides a framework for understanding each source of randomness and is useful for modeling W.

EXAMPLE 3.61

It takes John an average of 18 minutes each day to commute to work. What would you expect his average commute time to be for the week?

E We were told that the average (i.e. expected value) of the commute time is 18 minutes per day: $E(X_i) = 18$. To get the expected time for the sum of the five days, we can add up the expected time for each individual day:

$$\begin{aligned} E(W) &= E(X_1 + X_2 + X_3 + X_4 + X_5) \\ &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= 18 + 18 + 18 + 18 + 18 = 90 \text{ minutes} \end{aligned}$$

The expectation of the total time is equal to the sum of the expected individual times. More generally, the expectation of a sum of random variables is always the sum of the expectation for each random variable.

GUIDED PRACTICE 3.62

G Elena is selling a TV at a cash auction and also intends to buy a toaster oven in the auction. If X represents the profit for selling the TV and Y represents the cost of the toaster oven, write an equation that represents the net change in Elena's cash.[43]

GUIDED PRACTICE 3.63

G Based on past auctions, Elena figures she should expect to make about \$175 on the TV and pay about \$23 for the toaster oven. In total, how much should she expect to make or spend?[44]

i (scenario)	1 (`noBook`)	2 (`textbook`)	3 (`both`)	Total
y_i	0.00	159.00	200.00	
$P(Y = y_i)$	0.15	0.25	0.60	
$y_i \times P(Y = y_i)$	0.00	39.75	120.00	$E(Y) = 159.75$
$y_i - E(Y)$	-159.75	-0.75	40.25	
$(y_i - E(Y))^2$	25520.06	0.56	1620.06	
$(y_i - E(Y))^2 \times P(Y)$	3828.0	0.1	972.0	$Var(Y) \approx 4800$

[43] She will make X dollars on the TV but spend Y dollars on the toaster oven: $X - Y$.
[44] $E(X - Y) = E(X) - E(Y) = 175 - 23 = \$152$. She should expect to make about \$152.

GUIDED PRACTICE 3.64

G Would you be surprised if John's weekly commute wasn't exactly 90 minutes or if Elena didn't make exactly \$152? Explain.[45]

Two important concepts concerning combinations of random variables have so far been introduced. First, a final value can sometimes be described as the sum of its parts in an equation. Second, intuition suggests that putting the individual average values into this equation gives the average value we would expect in total. This second point needs clarification – it is guaranteed to be true in what are called *linear combinations of random variables*.

A **linear combination** of two random variables X and Y is a fancy phrase to describe a combination

$$aX + bY$$

where a and b are some fixed and known numbers. For John's commute time, there were five random variables – one for each work day – and each random variable could be written as having a fixed coefficient of 1:

$$1X_1 + 1X_2 + 1X_3 + 1X_4 + 1X_5$$

For Elena's net gain or loss, the X random variable had a coefficient of +1 and the Y random variable had a coefficient of -1.

When considering the average of a linear combination of random variables, it is safe to plug in the mean of each random variable and then compute the final result. For a few examples of nonlinear combinations of random variables – cases where we cannot simply plug in the means – see the footnote.[46]

LINEAR COMBINATIONS OF RANDOM VARIABLES AND THE AVERAGE RESULT

If X and Y are random variables, then a linear combination of the random variables is given by

$$aX + bY$$

where a and b are some fixed numbers. To compute the average value of a linear combination of random variables, plug in the average of each individual random variable and compute the result:

$$a \times E(X) + b \times E(Y)$$

Recall that the expected value is the same as the mean, e.g. $E(X) = \mu_X$.

[The next example about stock returns have been updated and continues into the text of Section 3.4.4.]

[45] No, since there is probably some variability. For example, the traffic will vary from one day to next, and auction prices will vary depending on the quality of the merchandise and the interest of the attendees.

[46] If X and Y are random variables, consider the following combinations: X^{1+Y}, $X \times Y$, X/Y. In such cases, plugging in the average value for each random variable and computing the result will not generally lead to an accurate average value for the end result.

EXAMPLE 3.65

Leonard has invested $6000 in Caterpillar Inc (stock ticker: CAT) and $2000 in Exxon Mobil Corp (XOM). If X represents the change in Caterpillar's stock next month and Y represents the change in Exxon Mobil's stock next month, write an equation that describes how much money will be made or lost in Leonard's stocks for the month.

(E) For simplicity, we will suppose X and Y are not in percents but are in decimal form (e.g. if Caterpillar's stock increases 1%, then $X = 0.01$; or if it loses 1%, then $X = -0.01$). Then we can write an equation for Leonard's gain as

$$\$6000 \times X + \$2000 \times Y$$

If we plug in the change in the stock value for X and Y, this equation gives the change in value of Leonard's stock portfolio for the month. A positive value represents a gain, and a negative value represents a loss.

GUIDED PRACTICE 3.66

(G) Caterpillar stock has recently been rising at 2.0% and Exxon Mobil's at 0.2% per month, respectively. Compute the expected change in Leonard's stock portfolio for next month.[47]

GUIDED PRACTICE 3.67

(G) You should have found that Leonard expects a positive gain in Guided Practice 3.66. However, would you be surprised if he actually had a loss this month?[48]

3.4.4 Variability in linear combinations of random variables

Quantifying the average outcome from a linear combination of random variables is helpful, but it is also important to have some sense of the uncertainty associated with the total outcome of that combination of random variables. The expected net gain or loss of Leonard's stock portfolio was considered in Guided Practice 3.66. However, there was no quantitative discussion of the volatility of this portfolio. For instance, while the average monthly gain might be about $124 according to the data, that gain is not guaranteed. Figure 3.24 shows the monthly changes in a portfolio like Leonard's during a three year period. The gains and losses vary widely, and quantifying these fluctuations is important when investing in stocks.

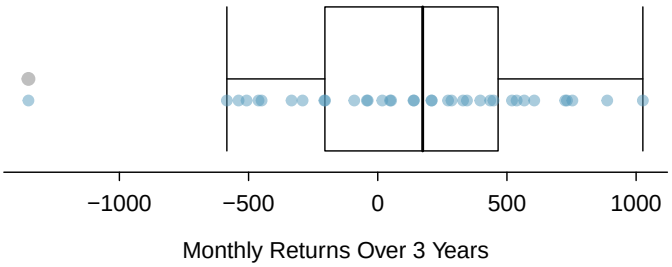


Figure 3.24: The change in a portfolio like Leonard's for 36 months, where $6000 is in Caterpillar's stock and $2000 is in Exxon Mobil's.

Just as we have done in many previous cases, we use the variance and standard deviation to describe the uncertainty associated with Leonard's monthly returns. To do so, the variances of each

[47] $E(\$6000 \times X + \$2000 \times Y) = \$6000 \times 0.020 + \$2000 \times 0.002 = \$124$.
[48] No. While stocks tend to rise over time, they are often volatile in the short term.

	Mean ($\bar{x}$)	Standard deviation (s)	Variance (s^2)
CAT	0.0204	0.0757	0.0057
XOM	0.0025	0.0455	0.0021

Figure 3.25: The mean, standard deviation, and variance of the CAT and XOM stocks. These statistics were estimated from historical stock data, so notation used for sample statistics has been used.

stock's monthly return will be useful, and these are shown in Figure 3.25. The stocks' returns are nearly independent.

Here we use an equation from probability theory to describe the uncertainty of Leonard's monthly returns; we leave the proof of this method to a dedicated probability course. The variance of a linear combination of random variables can be computed by plugging in the variances of the individual random variables and squaring the coefficients of the random variables:

$$Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

It is important to note that this equality assumes the random variables are independent; [new description here about if independence is broken] if independence doesn't hold, then a modification to this equation would be required that we leave as a topic for a future course to cover. This equation can be used to compute the variance of Leonard's monthly return:

$$\begin{aligned} Var(6000 \times X + 2000 \times Y) &= 6000^2 \times Var(X) + 2000^2 \times Var(Y) \\ &= 36,000,000 \times 0.0057 + 4,000,000 \times 0.0021 \\ &\approx 213,600 \end{aligned}$$

The standard deviation is computed as the square root of the variance: $\sqrt{213,600} = \$463$. While an average monthly return of \$124 on an \$8000 investment is nothing to scoff at, the monthly returns are so volatile that Leonard should not expect this income to be very stable.

VARIABILITY OF LINEAR COMBINATIONS OF RANDOM VARIABLES

The variance of a linear combination of random variables may be computed by squaring the constants, substituting in the variances for the random variables, and computing the result:

$$Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

This equation is valid as long as the random variables are independent of each other. The standard deviation of the linear combination may be found by taking the square root of the variance.

EXAMPLE 3.68

Suppose John's daily commute has a standard deviation of 4 minutes. What is the uncertainty in his total commute time for the week?

The expression for John's commute time was

$$X_1 + X_2 + X_3 + X_4 + X_5$$

Each coefficient is 1, and the variance of each day's time is $4^2 = 16$. Thus, the variance of the total weekly commute time is

$$\begin{aligned} \text{variance} &= 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 = 5 \times 16 = 80 \\ \text{standard deviation} &= \sqrt{\text{variance}} = \sqrt{80} = 8.94 \end{aligned}$$

The standard deviation for John's weekly work commute time is about 9 minutes.

GUIDED PRACTICE 3.69

G The computation in Example 3.68 relied on an important assumption: the commute time for each day is independent of the time on other days of that week. Do you think this is valid? Explain.[49]

GUIDED PRACTICE 3.70

G Consider Elena's two auctions from Guided Practice 3.62 on page 89. Suppose these auctions are approximately independent and the variability in auction prices associated with the TV and toaster oven can be described using standard deviations of $25 and $8. Compute the standard deviation of Elena's net gain.[50]

Consider again Guided Practice 3.70. The negative coefficient for Y in the linear combination was eliminated when we squared the coefficients. This generally holds true: negatives in a linear combination will have no impact on the variability computed for a linear combination, but they do impact the expected value computations.

[49] One concern is whether traffic patterns tend to have a weekly cycle (e.g. Fridays may be worse than other days). If that is the case, and John drives, then the assumption is probably not reasonable. However, if John walks to work, then his commute is probably not affected by any weekly traffic cycle.

[50] The equation for Elena can be written as

$$(1) \times X + (-1) \times Y$$

The variances of X and Y are 625 and 64. We square the coefficients and plug in the variances:

$$(1)^2 \times Var(X) + (-1)^2 \times Var(Y) = 1 \times 625 + 1 \times 64 = 689$$

The variance of the linear combination is 689, and the standard deviation is the square root of 689: about $26.25.

3.5 Continuous distributions

EXAMPLE 3.71

Figure 3.26 shows a few different hollow histograms for the heights of US adults. How does changing the number of bins allow you to make different interpretations of the data?

(E)

Adding more bins provides greater detail. This sample is extremely large, which is why much smaller bins still work well. Usually we do not use so many bins with smaller sample sizes since small counts per bin mean the bin heights are very volatile.

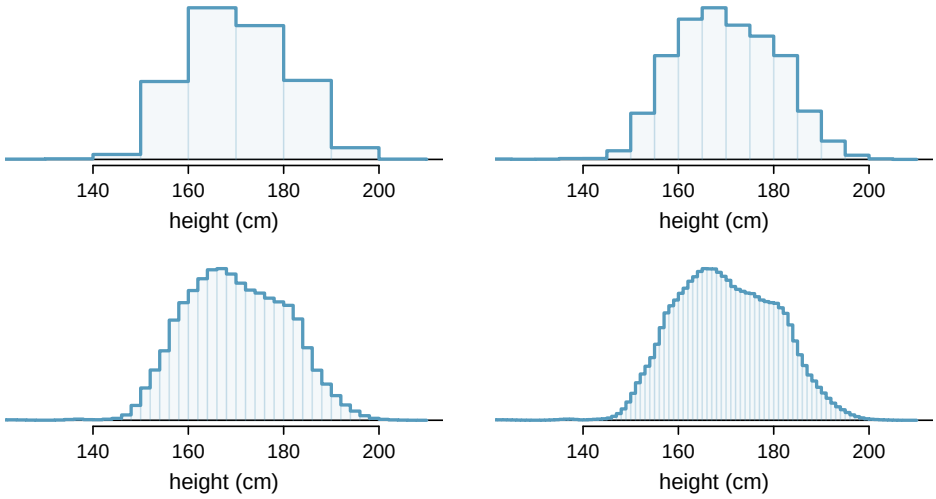


Figure 3.26: Four hollow histograms of US adults heights with varying bin widths.

EXAMPLE 3.72

What proportion of the sample is between `180` cm and `185` cm tall (about 5'11" to 6'1")?

(E)

We can add up the heights of the bins in the range `180` cm and `185` and divide by the sample size. For instance, this can be done with the two shaded bins shown in Figure 3.27. The two bins in this region have counts of 195,307 and 156,239 people, resulting in the following estimate of the probability:

$$\frac{195307 + 156239}{3{,}000{,}000} = 0.1172$$

This fraction is the same as the proportion of the histogram's area that falls in the range `180` to `185` cm.

3.5.1 From histograms to continuous distributions

Examine the transition from a boxy hollow histogram in the top-left of Figure 3.26 to the much smoother plot in the lower-right. In this last plot, the bins are so slim that the hollow histogram is starting to resemble a smooth curve. This suggests the population height as a *continuous* numerical variable might best be explained by a curve that represents the outline of extremely slim bins.

This smooth curve represents a **probability density function** (also called a **density** or **distribution**), and such a curve is shown in Figure 3.28 overlaid on a histogram of the sample. A

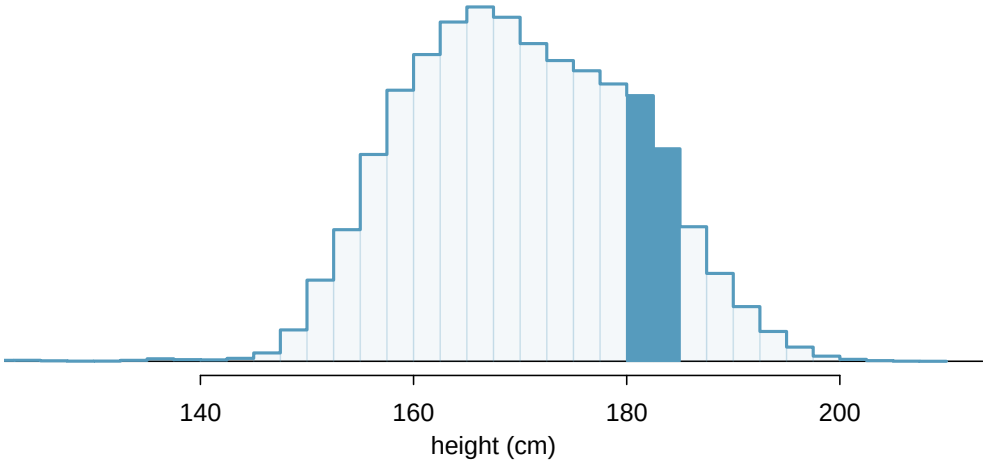


Figure 3.27: A histogram with bin sizes of 2.5 cm. The shaded region represents individuals with heights between 180 and 185 cm.

density has a special property: the total area under the density's curve is 1.

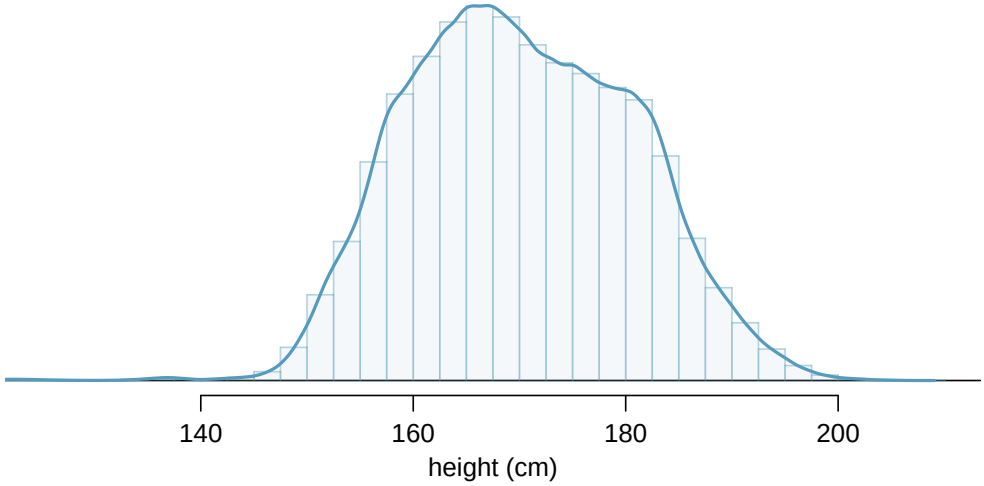


Figure 3.28: The continuous probability distribution of heights for US adults.

3.5.2 Probabilities from continuous distributions

We computed the proportion of individuals with heights 180 to 185 cm in Example 3.72 as a fraction:

$$\frac{\text{number of people between 180 and 185}}{\text{total sample size}}$$

We found the number of people with heights between 180 and 185 cm by determining the fraction of the histogram's area in this region. Similarly, we can use the area in the shaded region under the curve to find a probability (with the help of a computer):

$$P(\texttt{height} \text{ between } 180 \text{ and } 185) = \text{area between } 180 \text{ and } 185 = 0.1157$$

The probability that a randomly selected person is between 180 and 185 cm is 0.1157. This is very close to the estimate from Example 3.72: 0.1172.

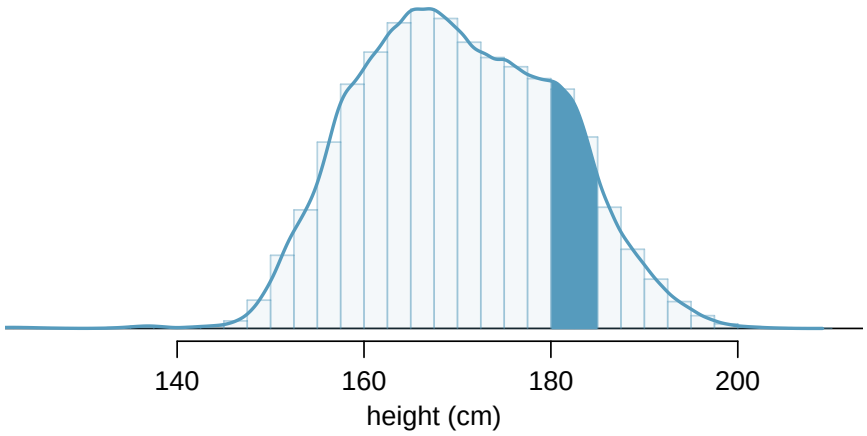


Figure 3.29: Density for heights in the US adult population with the area between 180 and 185 cm shaded. Compare this plot with Figure 3.27.

GUIDED PRACTICE 3.73

Three US adults are randomly selected. The probability a single adult is between 180 and 185 cm is 0.1157.[51]

(a) What is the probability that all three are between 180 and 185 cm tall?

(b) What is the probability that none are between 180 and 185 cm?

EXAMPLE 3.74

What is the probability that a randomly selected person is **exactly** 180 cm? Assume you can measure perfectly.

This probability is zero. A person might be close to 180 cm, but not exactly 180 cm tall. This also makes sense with the definition of probability as area; there is no area captured between 180 cm and 180 cm.

GUIDED PRACTICE 3.75

Suppose a person's height is rounded to the nearest centimeter. Is there a chance that a random person's **measured** height will be 180 cm?[52]

[51] Brief answers: (a) $0.1157 \times 0.1157 \times 0.1157 = 0.0015$. (b) $(1 - 0.1157)^3 = 0.692$

[52] This has positive probability. Anyone between 179.5 cm and 180.5 cm will have a *measured* height of 180 cm. This is probably a more realistic scenario to encounter in practice versus Example 3.74.

Chapter 4

Distributions of random variables

In this chapter, we discuss statistical distributions that frequently arise in the context of data analysis or statistical inference. We start with the normal distribution in the first section, which is used frequently in later chapters of this book. The remaining sections will occasionally be referenced but may be considered optional for the content in this book.

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

4.1 Normal distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**,[1] shown in Figure 4.1. Variables such as SAT scores and heights of US adult males closely follow the normal distribution.

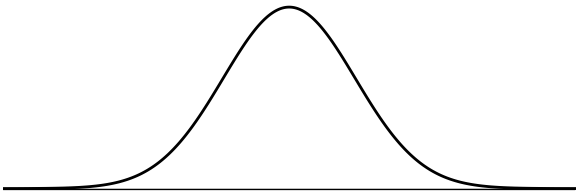


Figure 4.1: A normal curve.

NORMAL DISTRIBUTION FACTS

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

4.1.1 Normal distribution model

The normal distribution always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve. Figure 4.2 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure 4.3 shows these distributions on the same axis.

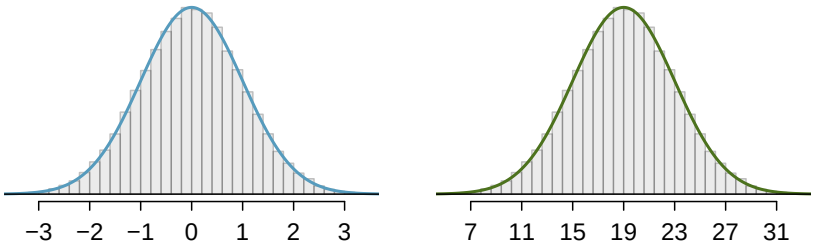


Figure 4.2: Both curves represent the normal distribution. However, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

If a normal distribution has mean μ and standard deviation σ, we may write the distribution as $N(\mu, \sigma)$. The two distributions in Figure 4.3 may be written as

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

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

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

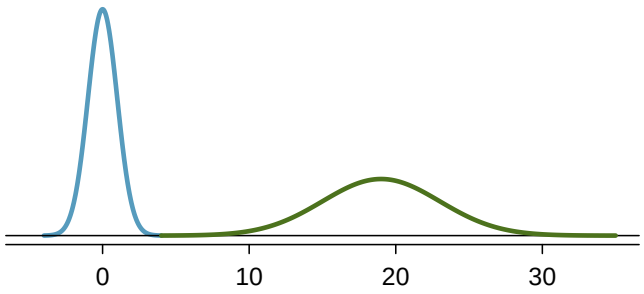


Figure 4.3: The normal models shown in Figure 4.2 but plotted together and on the same scale.

GUIDED PRACTICE 4.1

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

4.1.2 Standardizing with Z-scores

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

EXAMPLE 4.2

Table 4.4 shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT: $1100 + 200 = 1300$. Tom is 0.5 standard deviations above the mean on the ACT: $21 + 0.5 \times 6 = 24$. In Figure 4.5, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

	SAT	ACT
Mean	1100	21
SD	200	6

Figure 4.4: Mean and standard deviation for the SAT and ACT.

Example 4.2 used a standardization technique called a Z-score, a method most commonly employed for nearly normal observations but that may be used with any distribution. The **Z-score** of an observation is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its Z-score is 1. If it is 1.5 standard deviations *below* the mean, then its Z-score is -1.5. If x is an observation from a distribution $N(\mu, \sigma)$, we define the Z-score mathematically as

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

Using $\mu_{SAT} = 1100$, $\sigma_{SAT} = 200$, and $x_{\text{Ann}} = 1300$, we find Ann's Z-score:

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

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

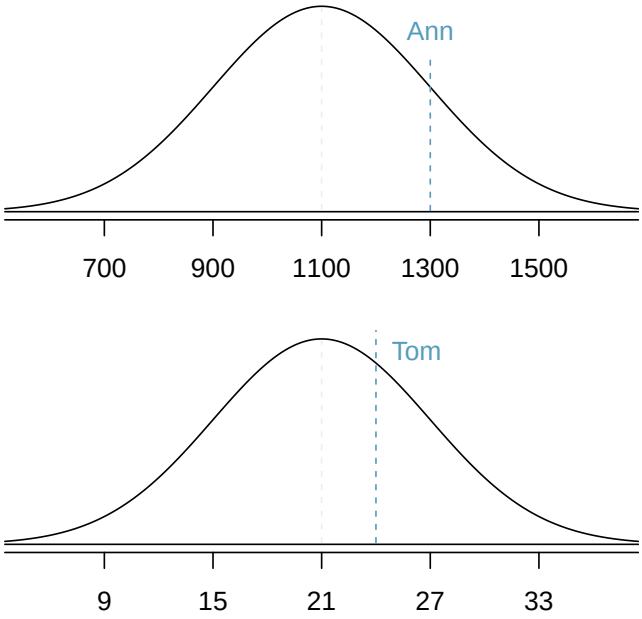


Figure 4.5: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

THE Z-SCORE

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

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

GUIDED PRACTICE 4.3

Use Tom's ACT score, 24, along with the ACT mean and standard deviation to compute his Z-score.[3]

Observations above the mean always have positive Z-scores, while those below the mean always have negative Z-scores. If an observation is equal to the mean, such as an SAT score of 1100, then the Z-score is 0.

GUIDED PRACTICE 4.4

Let X represent a random variable from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$.
(a) Find the Z-score of x.
(b) Use the Z-score to determine how many standard deviations above or below the mean x falls.[4]

GUIDED PRACTICE 4.5

Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm. Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm.[5]

[3] $Z_{Tom} = \frac{x_{\text{Tom}} - \mu_{\text{ACT}}}{\sigma_{\text{ACT}}} = \frac{24-21}{6} = 0.5$

[4] (a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$. (b) The observation x is 1.095 standard deviations *above* the mean. We know it must be above the mean since Z is positive.

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

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

G **GUIDED PRACTICE 4.6**

Which of the observations in Guided Practice 4.5 is more unusual?[6]

4.1.3 Finding tail areas

It's very useful in statistics to be able to identify tail areas of distributions. For instance, how many people have an SAT score below Ann's score of 1300? This is the same as Ann's **percentile**, which is the fraction of cases that have lower scores than Ann. We can visualize such a tail area like the curve and shading shown in Figure 4.6.

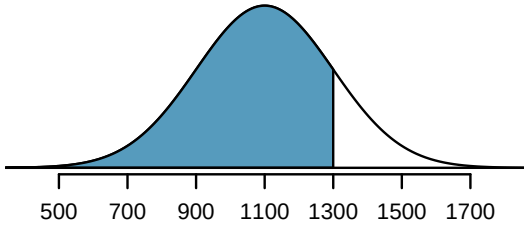


Figure 4.6: The area to the left of Z represents the percentile of the observation.

There are many techniques for doing this, and we'll discuss three of the options.

1. The most common approach in practice is to use statistical software. For example, in the program **R**, we could find the area shown in Figure 4.6 using the following command, which takes in the Z-score and returns the lower tail area:

   ```
   > pnorm(1)
   [1] 0.8413447
   ```

   According to this calculation, the region shaded that is below 1300 represents the proportion 0.841 (84.1%) of SAT test takers who had Z-scores below $Z = 1$. More generally, we can also specify the cutoff explicitly if we also note the mean and standard deviation:

   ```
   > pnorm(1300, mean = 1100, sd = 200)
   [1] 0.8413447
   ```

   There are many other software options, such as Python or SAS; even spreadsheet programs such as Excel and Google Sheets support these calculations.

2. A common strategy in classrooms is to use a graphing calculator, such as a TI or Casio calculator. These calculators require a series of button presses that are less concisely described. You can find instructions on using these calculators for finding tail areas of a normal distribution in the OpenIntro video library:

   www.openintro.org/videos

3. The last option for finding tail areas is to use what's called a **probability table**; these are occasionally used in classrooms but rarely in practice. Appendix B.1 contains such a table and a guide for how to use it.

We will solve normal distribution problems in this section by always first finding the Z-score. The reason is that we will encounter close parallels called test statistics beginning in Chapter 5; these are, in many instances, an equivalent of a Z-score.

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

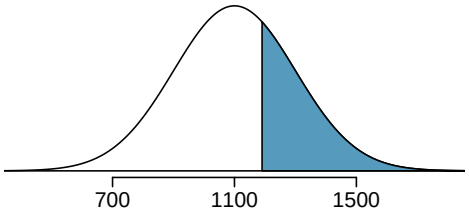
4.1.4 Normal probability examples

Cumulative SAT scores are approximated well by a normal model, $N(\mu = 1100, \sigma = 200)$.

EXAMPLE 4.7

Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1190 on her SATs?

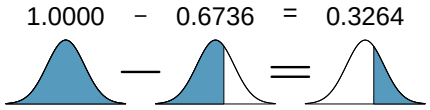
First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the chance she scores above 1190, so we shade this upper tail:



(E) The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z-score of the cutoff value. With $\mu = 1100$, $\sigma = 200$, and the cutoff value $x = 1190$, the Z-score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

Using statistical software (or another preferred method), we can area left of $Z = 0.45$ as 0.6736. To find the area *above* $Z = 0.45$, we compute one minus the area of the lower tail:



The probability Shannon scores at least 1190 on the SAT is 0.3264.

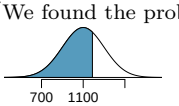
ALWAYS DRAW A PICTURE FIRST, AND FIND THE Z-SCORE SECOND

For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability. After drawing a figure to represent the situation, identify the Z-score for the value of interest.

GUIDED PRACTICE 4.8

(G) If the probability of Shannon scoring at least 1190 is 0.3264, then what is the probability she scores less than 1190? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[7]

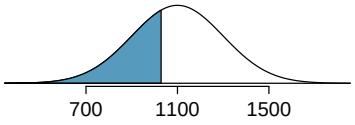
[7] We found the probability in Example 4.7: 0.6736.



EXAMPLE 4.9

Edward earned a 1030 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1030. These are the scores to the left of 1030.



Identifying the mean $\mu = 1100$, the standard deviation $\sigma = 200$, and the cutoff for the tail area $x = 1030$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Using statistical software, we get a tail area of 0.3632. Edward is at the 36^{th} percentile.

GUIDED PRACTICE 4.10

Use the results of Example 4.9 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[8]

FINDING AREAS TO THE RIGHT

Many software programs return the area to the left when given a Z-score. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

GUIDED PRACTICE 4.11

Stuart earned an SAT score of 1500. Draw a picture for each part.
(a) What is his percentile?
(b) What percent of SAT takers did better than Stuart?[9]

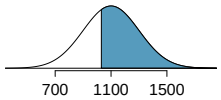
Based on a sample of 100 men, the heights of male adults in the US is nearly normal with mean 70.0" and standard deviation 3.3".

GUIDED PRACTICE 4.12

Mike is 5'7" and Jose is 6'4", and they both live in the US.
(a) What is Mike's height percentile?
(b) What is Jose's height percentile?
Also draw one picture for each part.[10]

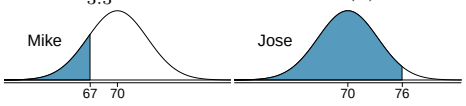
The last several problems have focused on finding the percentile (or upper tail) for a particular observation. What if you would like to know the observation corresponding to a particular percentile?

[8] If Edward did better than 36% of SAT takers, then about 64% must have done better than him.



[9] We leave the drawings to you. (a) $Z = \frac{1500-1100}{200} = 2 \rightarrow 0.9772$. (b) $1 - 0.9772 = 0.0228$.

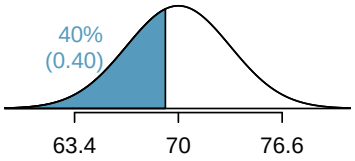
[10] First put the heights into inches: 67 and 76 inches. Figures are shown below.
(a) $Z_{\text{Mike}} = \frac{67-70}{3.3} = -0.91 \rightarrow 0.1814$. (b) $Z_{\text{Jose}} = \frac{76-70}{3.3} = 1.82 \rightarrow 0.9656$.



EXAMPLE 4.13

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

As always, first draw the picture.



(E)

In this case, the lower tail probability is known (0.40), which can be shaded on the diagram. We want to find the observation that corresponds to this value. As a first step in this direction, we determine the Z-score associated with the 40^{th} percentile. Using software, we can obtain the corresponding Z-score of about -0.25.

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

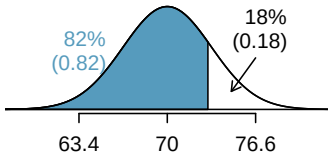
$$-0.25 = Z_{\text{Erik}} = \frac{x_{\text{Erik}} - \mu}{\sigma} = \frac{x_{\text{Erik}} - 70}{3.3}$$

Solving for x_{Erik} yields a height of 69.18 inches. That is, Erik is about 5'9".

EXAMPLE 4.14

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

Again, we draw the figure first.



(E)

Next, we want to find the Z-score at the 82^{nd} percentile, which will be a positive value and can be found using software as $Z = 0.92$. Finally, the height x is found using the Z-score formula with the known mean μ, standard deviation σ, and Z-score $Z = 0.92$:

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

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

GUIDED PRACTICE 4.15

(G) The SAT scores follow $N(1100, 200)$.[11]
(a) What is the 95^{th} percentile for SAT scores?
(b) What is the 97.5^{th} percentile for SAT scores?

GUIDED PRACTICE 4.16

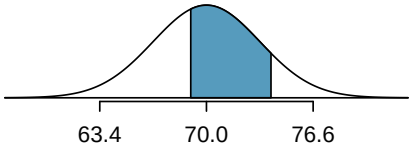
(G) Adult male heights follow $N(70.0", 3.3")$.[12]
(a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)?
(b) What is the probability that a male adult is shorter than 5'9" (69 inches)?

[11] Short answers: (a) $Z_{95} = 1.65 \rightarrow 1430$ SAT score. (b) $Z_{97.5} = 1.96 \rightarrow 1492$ SAT score.
[12] Short answers: (a) $Z = 1.21 \rightarrow 0.8869$, then subtract this value from 1 to get 0.1131. (b) $Z = -0.30 \rightarrow 0.3821$.

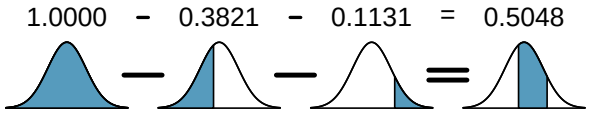
EXAMPLE 4.17

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

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



The total area under the curve is 1. If we find the area of the two tails that are not shaded (from Guided Practice 4.16, these areas are 0.3821 and 0.1131), then we can find the middle area:



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

GUIDED PRACTICE 4.18

SAT scores follow $N(1100, 200)$. What percent of SAT takers get between 1100 and 1400?[13]

GUIDED PRACTICE 4.19

Adult male heights follow $N(70.0", 3.3")$. What percent of adult males are between 5'5" and 5'7"?[14]

4.1.5 68-95-99.7 rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z-table.

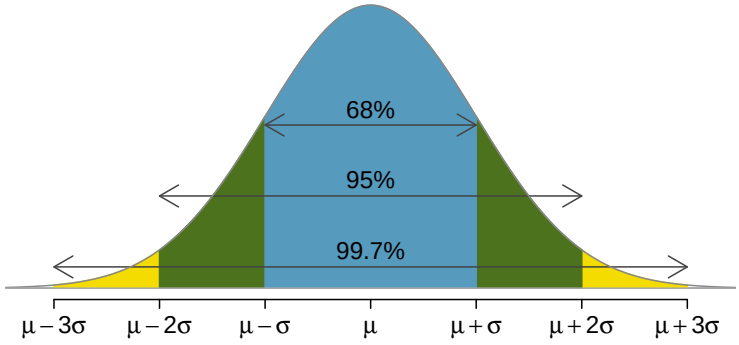


Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

[13] This is an abbreviated solution. (Be sure to draw a figure!) First find the percent who get below 1100 and the percent that get above 1400: $Z_{1100} = 0.00 \rightarrow 0.5000$ (area below), $Z_{1400} = 1.5 \rightarrow 0.0668$ (area above). Final answer: $1.0000 - 0.5000 - 0.0668 = 0.4332$.

[14] 5'5" is 65 inches ($Z = -1.52$). 5'7" is 67 inches ($Z = -0.91$). Numerical solution: $1.000 - 0.0643 - 0.8186 = 0.1171$, i.e. 11.71%.

GUIDED PRACTICE 4.20

Use software, a calculator, or a probability table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[15]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-15,000. For 5 and 6 standard deviations, it is about 1-in-2 million and 1-in-500 million, respectively.

GUIDED PRACTICE 4.21

SAT scores closely follow the normal model with mean $\mu = 1100$ and standard deviation $\sigma = 200$.[16]
(a) About what percent of test takers score 700 to 1500?
(b) What percent score between 1100 and 1500?

[15] First draw the pictures. Using software, we get 0.6827 within 1 standard deviation, 0.9545 within 2 standard deviations, and 0.9973 within 3 standard deviations.

[16] (a) 700 and 1500 represent two standard deviations below and above the mean, which means about 95% of test takers will score between 700 and 1500. (b) We found that 700 to 1500 represents about 95% of test takers. These test takers would be evenly split by the center of the distribution, 1100, so $\frac{95\%}{2} = 47.5\%$ of all test takers score between 1100 and 1500.

4.2 Geometric distribution

How long should we expect to flip a coin until it turns up `heads`? Or how many times should we expect to roll a die until we get a `1`? These questions can be answered using the geometric distribution. We first formalize each trial – such as a single coin flip or die toss – using the Bernoulli distribution, and then we combine these with our tools from probability (Chapter 3) to construct the geometric distribution.

4.2.1 Bernoulli distribution

Many health insurance plans in the United States have a deductible, where the insured individual is responsible for costs up to the deductible, and then the costs above the deductible are shared between the individual and insurance company for the remainder of the year.

Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year. Each of these people can be thought of as a **trial**. We label a person a **success** if her healthcare costs do not exceed the deductible. We label a person a **failure** if she does exceed her deductible in the year. Because 80% of the individuals will not hit their deductible, we denote the **probability of a success** as $p = 0.7$. The probability of a failure is sometimes denoted with $q = 1 - p$, which would be 0.3 in for the insurance example.

When an individual trial only has two possible outcomes, often labeled as `success` or `failure`, it is called a **Bernoulli random variable**. We chose to label a person who does not hit her deductible as a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

`1 1 1 0 1 0 0 1 1 0`

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable.[17]

BERNOULLI RANDOM VARIABLE

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

[17] If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned}\mu = E[X] &= P(X=0)\times 0 + P(X=1)\times 1 \\ &= (1-p)\times 0 + p\times 1 = 0 + p = p\end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned}\sigma^2 &= P(X=0)(0-p)^2 + P(X=1)(1-p)^2 \\ &= (1-p)p^2 + p(1-p)^2 = p(1-p)\end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

4.2.2 Geometric distribution

The **geometric distribution** is used to describe how many trials it takes to observe a success. Let's first look at an example.

EXAMPLE 4.22

Suppose we are working at the insurance company and need to find a case where the person did not exceed her (or his) deductible as a case study. If the probability a person will not exceed her deductible is 0.7 and we are drawing people at random, what are the chances that the first person will not have exceeded her deductible, i.e. be a success? The second person? The third? What about we pull $n-1$ cases before we find the first success, i.e. the first success is the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

(E) The probability of stopping after the first person is just the chance the first person will not hit her (or his) deductible: 0.7. The probability it will be the second person is

$$P(\text{second person is the first to not administer the worst shock})$$
$$= P(\text{the first will, the second won't}) = (0.3)(0.7) = 0.21$$

Likewise, the probability it will be the third case is $(0.3)(0.3)(0.7) = 0.063$.

If the first success is on the n^{th} person, then there are $n-1$ failures and finally 1 success, which corresponds to the probability $(0.3)^{n-1}(0.7)$. This is the same as $(1-0.7)^{n-1}(0.7)$.

Example 4.22 illustrates what the **geometric distribution**, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 4.22 is shown in Figure 4.8. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

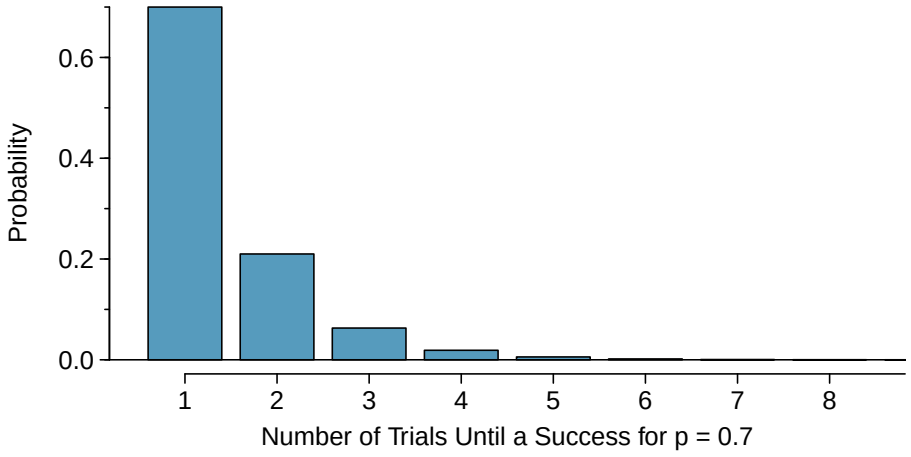


Figure 4.8: The geometric distribution when the probability of success is $p = 0.7$.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation or variance of this distribution, we present general formulas for each.

GEOMETRIC DISTRIBUTION

If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

It takes, on average, $1/p$ trials to get a success under the geometric distribution. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

GUIDED PRACTICE 4.23

(G) The probability that a particular case would not exceed their deductible is said to be 0.7. If we were to examine cases until we found one that where the person did not hit her deductible, how many cases should we expect to check?[18]

EXAMPLE 4.24

What is the chance that we would find the first success within the first 3 cases?

This is the chance it is the first ($n = 1$), second ($n = 2$), or third ($n = 3$) case is the first success, which are three disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

(E)

$$\begin{aligned} &P(n = 1, 2, \text{ or } 3) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) \\ &\quad = (0.3)^{1-1}(0.7) + (0.3)^{2-1}(0.7) + (0.3)^{3-1}(0.7) \\ &\quad = 0.973 \end{aligned}$$

There is a probability of 0.973 that we would find a successful case within 3 cases.

GUIDED PRACTICE 4.25

(G) Determine a more clever way to solve Example 4.24. Show that you get the same result.[19]

[18] We would expect to see about $1/0.7 \approx 1.43$ individuals to find the first success.

[19] First find the probability of the complement: $P(\text{no success in first 3 trials}) = 0.3^3 = 0.027$. Next, compute one minus this probability: $1 - P(\text{no success in 3 trials}) = 1 - 0.027 = 0.973$.

EXAMPLE 4.26

Suppose a car insurer has determined that 88% of its drivers will not exceed their deductible in a given year. If someone at the company were to randomly draw driver files until they found one that had not exceeded their deductible, what is the expected number of drivers the insurance employee must check? What is the standard deviation of the number of driver files that must be drawn?

(E)

In this example, a success is again when someone will not exceed the insurance deductible, which has probability $p = 0.88$. The expected number of people to be checked is $1/p = 1/0.88 = 1.14$ and the standard deviation is $\sqrt{(1-p)/p^2} = 0.39$.

GUIDED PRACTICE 4.27

(G) Using the results from Example 4.26, $\mu = 1.14$ and $\sigma = 0.39$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[20]

The independence assumption is crucial to the geometric distribution's accurate description of a scenario. Mathematically, we can see that to construct the probability of the success on the n^{th} trial, we had to use the Multiplication Rule for Independent Processes. It is no simple task to generalize the geometric model for dependent trials.

[20] No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.

4.3 Binomial distribution

The **binomial distribution** is used to describe the number of successes in a fixed number of trials. This is different from the geometric distribution, which described the number of trials we must wait before we observe a success.

4.3.1 The binomial distribution

Let's again imagine ourselves back at the insurance agency where 70% of individuals do not exceed their deductible.

EXAMPLE 4.28

Suppose the insurance agency is considering a random sample of four individuals they insure. What is the chance exactly one of them will exceed the deductible and the other four will not? Let's call the four people Ariana (A), Brittany (B), Carlton (C), and Damian (D) for convenience.

Let's consider a scenario where one person exceeds the deductible:

E

$$\begin{aligned} &P(A = \texttt{exceed},\ B = \texttt{not},\ C = \texttt{not},\ D = \texttt{not}) \\ &\quad = P(A = \texttt{exceed})\ P(B = \texttt{not})\ P(C = \texttt{not})\ P(D = \texttt{not}) \\ &\quad = (0.3)(0.7)(0.7)(0.7) \\ &\quad = (0.7)^3(0.3)^1 \\ &\quad = 0.103 \end{aligned}$$

But there are three other scenarios: Brittany, Carlton, or Damian could have been the one to exceed the deductible. In each of these cases, the probability is again $(0.7)^3(0.3)^1$. These four scenarios exhaust all the possible ways that exactly one of these four people could have exceeded the deductible, so the total probability is $4 \times (0.7)^3(0.3)^1 = 0.412$.

GUIDED PRACTICE 4.29

G Verify that the scenario where Brittany is the only one exceed the deductible has probability $(0.7)^3(0.3)^1$. [21]

The scenario outlined in Example 4.28 is an example of a binomial distribution scenario. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p (in Example 4.28, $n = 4$, $k = 3$, $p = 0.7$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, k, and p to obtain the probability. To do this, we reexamine each part of Example 4.28.

There were four individuals who could have been the one to exceed the deductible, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\#\text{ of scenarios}] \times P(\text{single scenario})$$

The first component of this equation is the number of ways to arrange the $k = 3$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider $P(\text{single scenario})$ under the general case of k successes and $n - k$ failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^k(1-p)^{n-k}$$

This is our general formula for $P(\text{single scenario})$.

[21] $P(A = \texttt{not},\ B = \texttt{exceed},\ C = \texttt{not},\ D = \texttt{not}) = (0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^1$.

Secondly, we introduce a general formula for the number of ways to choose k successes in n trials, i.e. arrange k successes and $n-k$ failures:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The quantity $\binom{n}{k}$ is read **n choose k**.[22] The exclamation point notation (e.g. $k!$) denotes a **factorial** expression.

$$\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \times 1 = 2 \\
3! &= 3 \times 2 \times 1 = 6 \\
4! &= 4 \times 3 \times 2 \times 1 = 24 \\
&\vdots \\
n! &= n \times (n-1) \times ... \times 3 \times 2 \times 1
\end{aligned}$$

Using the formula, we can compute the number of ways to choose $k = 3$ successes in $n = 4$ trials:

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!1!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(1)} = 4$$

This result is exactly what we found by carefully thinking of each possible scenario in Example 4.28.

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability yields the general binomial formula.

BINOMIAL DISTRIBUTION

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

The mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)}$$

IS IT BINOMIAL? FOUR CONDITIONS TO CHECK.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

[22] Other notation for n choose k includes ${}_nC_k$, C_n^k, and $C(n,k)$.

EXAMPLE 4.30

What is the probability that 3 of 8 randomly selected individuals will have exceeded the insurance deductible, i.e. that 5 of 8 will not exceed the deductible? Recall that 70% of individuals will not exceed the deductible.

We would like to apply the binomial model, so we check the conditions. The number of trials is fixed ($n = 8$) (condition 2) and each trial outcome can be classified as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are $k = 5$ successes in $n = 8$ trials (recall that a success is an individual who does *not* exceed the deductible, and the probability of a success is $p = 0.7$. So the probability that 5 of 8 will not exceed the deductible and 3 will exceed the deductible is given by

$$\binom{8}{5}(0.7)^5(1-0.7)^{8-5} = \frac{8!}{5!(5-3)!}(0.7)^5(1-0.7)^{8-5}$$
$$= \frac{8!}{5!3!}(0.7)^5(0.3)^3$$

Dealing with the factorial part:

$$\frac{8!}{5!3!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(5\times4\times3\times2\times1)(3\times2\times1)} = \frac{8\times7\times6}{3\times2\times1} = 56$$

Using $(0.7)^5(0.3)^3 \approx 0.00454$, the final probability is about $56 * 0.00454 \approx 0.254$.

COMPUTING BINOMIAL PROBABILITIES

The first step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and k. As the last stage use software or the formulas to determine the probability, then interpret the results.

If you must do calculations by hand, it's often useful to cancel out as many terms as possible in the top and bottom of the binomial coefficient.

GUIDED PRACTICE 4.31

If we randomly sampled 40 case files from the insurance agency discussed earlier, how many of the cases would you expect to not have exceeded the deductible in a given year? What is the standard deviation of the number that would not have exceeded the deductible?[23]

GUIDED PRACTICE 4.32

The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3. If you have 4 friends who smoke, are the conditions for the binomial model satisfied?[24]

---

[23] We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas: $\mu = np = 40 \times 0.7 = 28$ and $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.7 \times 0.3} = 2.9$. Because very roughly 95% of observations fall within 2 standard deviations of the mean (see Section 2.1.4), we would probably observe at least 22 but fewer than 34 individuals in our sample who would not exceed the deductible.

[24] One possible answer: if the friends know each other, then the independence assumption is probably not satisfied. For example, acquaintances may have similar smoking habits, or those friends might make a pact to quit together.

GUIDED PRACTICE 4.33

Suppose these four friends do not know each other and we can treat them as if they were a random sample from the population. Is the binomial model appropriate? What is the probability that[25]

(a) None of them will develop a severe lung condition?

(b) One will develop a severe lung condition?

(c) That no more than one will develop a severe lung condition?

GUIDED PRACTICE 4.34

What is the probability that at least 2 of your 4 smoking friends will develop a severe lung condition in their lifetimes?[26]

GUIDED PRACTICE 4.35

Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers.[27]

(a) How many would you expect to develop a severe lung condition, i.e. what is the mean?

(b) What is the probability that at most 2 of your 7 friends will develop a severe lung condition.

Next we consider the first term in the binomial probability, n choose k under some special scenarios.

GUIDED PRACTICE 4.36

Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[28]

GUIDED PRACTICE 4.37

How many ways can you arrange one success and $n-1$ failures in n trials? How many ways can you arrange $n-1$ successes and one failure in n trials?[29]

[25] To check if the binomial model is appropriate, we must verify the conditions. (i) Since we are supposing we can treat the friends as a random sample, they are independent. (ii) We have a fixed number of trials ($n = 4$). (iii) Each outcome is a success or failure. (iv) The probability of a success is the same for each trials since the individuals are like a random sample ($p = 0.3$ if we say a "success" is someone getting a lung condition, a morbid choice). Compute parts (a) and (b) using the binomial formula: $P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$, $P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116$. Note: $0! = 1$. Part (c) can be computed as the sum of parts (a) and (b): $P(0) + P(1) = 0.2401 + 0.4116 = 0.6517$. That is, there is about a 65% chance that no more than one of your four smoking friends will develop a severe lung condition.

[26] The complement (no more than one will develop a severe lung condition) as computed in Guided Practice 4.33 as 0.6517, so we compute one minus this value: 0.3483.

[27] (a) $\mu = 0.3 \times 7 = 2.1$. (b) $P(0, 1, \text{or } 2 \text{ develop severe lung condition}) = P(k = 0) + P(k = 1) + P(k = 2) = 0.6471$.

[28] Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

[29] One success and $n-1$ failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and $n-1$ failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

4.3.2 Normal approximation to the binomial distribution

The binomial formula is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities.

EXAMPLE 4.38

Approximately 15% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?

We leave the usual verification that the four conditions for the binomial model are valid as an exercise.

(E) The question posed is equivalent to asking, what is the probability of observing $k = 0$, 1, 2, ..., or 42 smokers in a sample of $n = 400$ when $p = 0.15$? We can compute these 43 different probabilities and add them together to find the answer:

$$\begin{aligned} &P(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 42) \\ &\quad = P(k = 0) + P(k = 1) + \cdots + P(k = 42) \\ &\quad = 0.0054 \end{aligned}$$

If the true proportion of smokers in the community is $p = 0.15$, then the probability of observing 42 or fewer smokers in a sample of $n = 400$ is 0.0054.

The computations in Example 4.38 are tedious and long. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. We might wonder, is it reasonable to use the normal model in place of the binomial distribution? Surprisingly, yes, if certain conditions are met.

GUIDED PRACTICE 4.39

(G) Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 4.9 shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10, 30, 100, 300$. What happens to the shape of the distributions as the sample size increases? What distribution does the last hollow histogram resemble?[30]

NORMAL APPROXIMATION OF THE BINOMIAL DISTRIBUTION

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1 - p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of Example 4.38.

[30] The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in last hollow histogram.

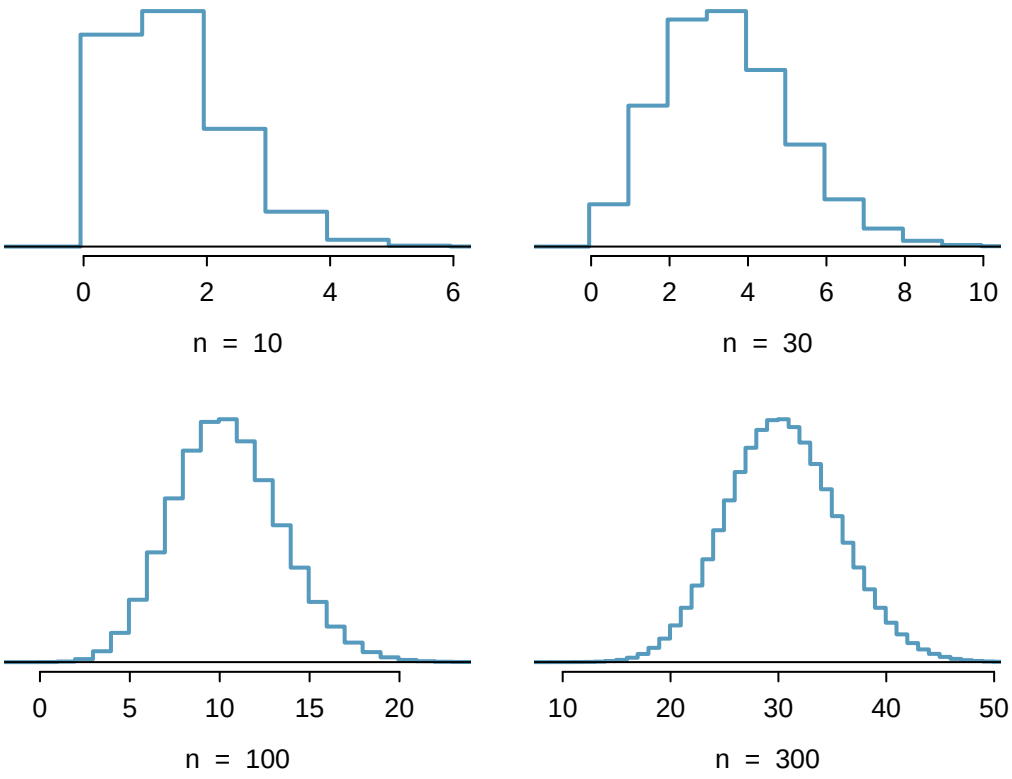


Figure 4.9: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

EXAMPLE 4.40

How can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

Showing that the binomial model is reasonable was a suggested exercise in Example 4.38. We also verify that both np and $n(1-p)$ are at least 10:

(E)

$$np = 400 \times 0.15 = 60 \qquad n(1-p) = 400 \times 0.85 = 340$$

With these conditions checked, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

$$\mu = np = 60 \qquad \sigma = \sqrt{np(1-p)} = 7.14$$

We want to find the probability of observing 42 or fewer smokers using this model.

GUIDED PRACTICE 4.41

(G) Use the normal model $N(\mu = 60, \sigma = 7.14)$ to estimate the probability of observing 42 or fewer smokers. Your answer should be approximately equal to the solution of Example 4.38: 0.0054. [31]

[31] Compute the Z-score first: $Z = \frac{42-60}{7.14} = -2.52$. The corresponding left tail area is 0.0059.

4.3.3 The normal approximation breaks down on small intervals

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$. With such a large sample, we might be tempted to apply the normal approximation and use the range 49 to 51. However, we would find that the binomial solution and the normal approximation notably differ:

Binomial: 0.0649 Normal: 0.0421

We can identify the cause of this discrepancy using Figure 4.10, which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval.

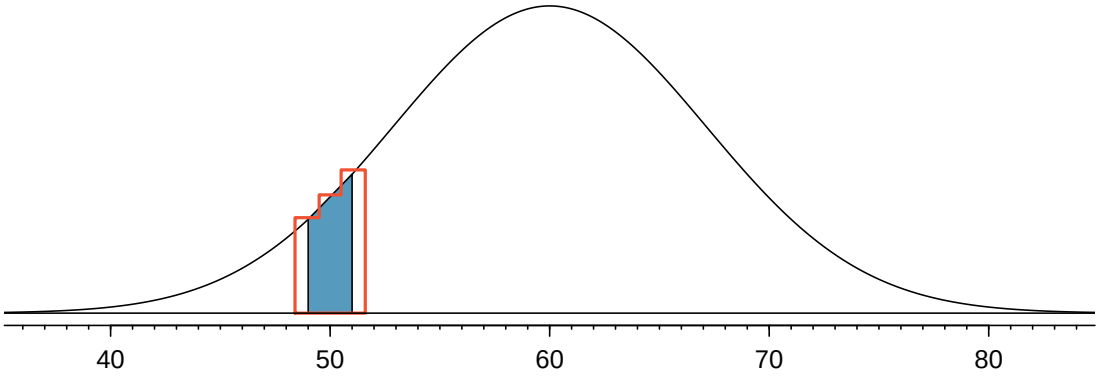


Figure 4.10: A normal curve with the area between 49 and 51 shaded. The outlined area represents the exact binomial probability.

IMPROVING THE NORMAL APPROXIMATION FOR THE BINOMIAL DISTRIBUTION

The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values are modified slightly. The cutoff values for the lower end of a shaded region should be reduced by 0.5, and the cutoff value for the upper end should be increased by 0.5.

The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. In the example above, the revised normal distribution estimate is 0.0633, much closer to the exact value of 0.0649. While it is possible to also apply this correction when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

4.4 Negative binomial distribution

The geometric distribution describes the probability of observing the first success on the n^{th} trial. The **negative binomial distribution** is more general: it describes the probability of observing the k^{th} success on the n^{th} trial.

EXAMPLE 4.42

Each day a high school football coach tells his star kicker, Brian, that he can go home after he successfully kicks four 35 yard field goals. Suppose we say each kick has a probability p of being successful. If p is small – e.g. close to 0.1 – would we expect Brian to need many attempts before he successfully kicks his fourth field goal?

(E)

We are waiting for the fourth success ($k = 4$). If the probability of a success (p) is small, then the number of attempts (n) will probably be large. This means that Brian is more likely to need many attempts before he gets $k = 4$ successes. To put this another way, the probability of n being small is low.

To identify a negative binomial case, we check 4 conditions. The first three are common to the binomial distribution.

IS IT NEGATIVE BINOMIAL? FOUR CONDITIONS TO CHECK

(1) The trials are independent.
(2) Each trial outcome can be classified as a success or failure.
(3) The probability of a success (p) is the same for each trial.
(4) The last trial must be a success.

GUIDED PRACTICE 4.43

(G) Suppose Brian is very diligent in his attempts and he makes each 35 yard field goal with probability $p = 0.8$. Take a guess at how many attempts he would need before making his fourth kick.[32]

EXAMPLE 4.44

In yesterday's practice, it took Brian only 6 tries to get his fourth field goal. Write out each of the possible sequence of kicks.

(E)

Because it took Brian six tries to get the fourth success, we know the last kick must have been a success. That leaves three successful kicks and two unsuccessful kicks (we label these as failures) that make up the first five attempts. There are ten possible sequences of these first five kicks, which are shown in Figure 4.11. If Brian achieved his fourth success ($k = 4$) on his sixth attempt ($n = 6$), then his order of successes and failures must be one of these ten possible sequences.

GUIDED PRACTICE 4.45

(G) Each sequence in Figure 4.11 has exactly two failures and four successes with the last attempt always being a success. If the probability of a success is $p = 0.8$, find the probability of the first sequence.[33]

If the probability Brian kicks a 35 yard field goal is $p = 0.8$, what is the probability it takes

[32] One possible answer: since he is likely to make each field goal attempt, it will take him at least 4 attempts but probably not more than 6 or 7.

[33] The first sequence: $0.2 \times 0.2 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.0164$.

	Kick Attempt					
	1	2	3	4	5	6
1	F	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
2	F	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
3	F	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
4	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
5	$\overset{1}{S}$	F	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
6	$\overset{1}{S}$	F	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
7	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
8	$\overset{1}{S}$	$\overset{2}{S}$	F	F	$\overset{3}{S}$	$\overset{4}{S}$
9	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	F	$\overset{4}{S}$
10	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	F	$\overset{4}{S}$

Figure 4.11: The ten possible sequences when the fourth successful kick is on the sixth attempt.

Brian exactly six tries to get his fourth successful kick? We can write this as

$$\begin{aligned} &P(\text{it takes Brian six tries to make four field goals}) \\ &\quad = P(\text{Brian makes three of his first five field goals, and he makes the sixth one}) \\ &\quad = P(1^{st} \text{ sequence OR } 2^{nd} \text{ sequence OR ... OR } 10^{th} \text{ sequence}) \end{aligned}$$

where the sequences are from Figure 4.11. We can break down this last probability into the sum of ten disjoint possibilities:

$$\begin{aligned} &P(1^{st} \text{ sequence OR } 2^{nd} \text{ sequence OR ... OR } 10^{th} \text{ sequence}) \\ &\quad = P(1^{st} \text{ sequence}) + P(2^{nd} \text{ sequence}) + \cdots + P(10^{th} \text{ sequence}) \end{aligned}$$

The probability of the first sequence was identified in Guided Practice 4.45 as 0.0164, and each of the other sequences have the same probability. Since each of the ten sequence has the same probability, the total probability is ten times that of any individual sequence.

The way to compute this negative binomial probability is similar to how the binomial problems were solved in Section 4.3. The probability is broken into two pieces:

$$\begin{aligned} &P(\text{it takes Brian six tries to make four field goals}) \\ &= [\text{Number of possible sequences}] \times P(\text{Single sequence}) \end{aligned}$$

Each part is examined separately, then we multiply to get the final result.

We first identify the probability of a single sequence. One particular case is to first observe all the failures ($n-k$ of them) followed by the k successes:

$$\begin{aligned} &P(\text{Single sequence}) \\ &= P(n-k \text{ failures and then } k \text{ successes}) \\ &= (1-p)^{n-k}p^k \end{aligned}$$

We must also identify the number of sequences for the general case. Above, ten sequences were identified where the fourth success came on the sixth attempt. These sequences were identified by fixing the last observation as a success and looking for all the ways to arrange the other observations. In other words, how many ways could we arrange $k-1$ successes in $n-1$ trials? This can be found using the n choose k coefficient but for $n-1$ and $k-1$ instead:

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)!\,((n-1)-(k-1))!} = \frac{(n-1)!}{(k-1)!\,(n-k)!}$$

This is the number of different ways we can order $k-1$ successes and $n-k$ failures in $n-1$ trials. If the factorial notation (the exclamation point) is unfamiliar, see page **??**.

NEGATIVE BINOMIAL DISTRIBUTION

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial, where all trials are independent:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

The value p represents the probability that an individual trial is a success.

EXAMPLE 4.46

Show using the formula for the negative binomial distribution that the probability Brian kicks his fourth successful field goal on the sixth attempt is 0.164.

(E) The probability of a single success is $p = 0.8$, the number of successes is $k = 4$, and the number of necessary attempts under this scenario is $n = 6$.

$$\binom{n-1}{k-1} p^k (1-p)^{n-k} \;=\; \frac{5!}{3!2!}(0.8)^4(0.2)^2 \;=\; 10 \times 0.0164 \;=\; 0.164$$

GUIDED PRACTICE 4.47

(G) The negative binomial distribution requires that each kick attempt by Brian is independent. Do you think it is reasonable to suggest that each of Brian's kick attempts are independent?[34]

GUIDED PRACTICE 4.48

(G) Assume Brian's kick attempts are independent. What is the probability that Brian will kick his fourth field goal within 5 attempts?[35]

BINOMIAL VERSUS NEGATIVE BINOMIAL

In the binomial case, we typically have a fixed number of trials and instead consider the number of successes. In the negative binomial case, we examine how many trials it takes to observe a fixed number of successes and require that the last observation be a success.

[34] Answers may vary. We cannot conclusively say they are or are not independent. However, many statistical reviews of athletic performance suggests such attempts are very nearly independent.

[35] If his fourth field goal ($k = 4$) is within five attempts, it either took him four or five tries ($n = 4$ or $n = 5$). We have $p = 0.8$ from earlier. Use the negative binomial distribution to compute the probability of $n = 4$ tries and $n = 5$ tries, then add those probabilities together:

$$\begin{aligned} &P(n = 4 \text{ OR } n = 5) = P(n = 4) + P(n = 5) \\ &= \binom{4-1}{4-1} 0.8^4 + \binom{5-1}{4-1}(0.8)^4(1-0.8) = 1 \times 0.41 + 4 \times 0.082 = 0.41 + 0.33 = 0.74 \end{aligned}$$

GUIDED PRACTICE 4.49

On 70% of days, a hospital admits at least one heart attack patient. On 30% of the days, no heart attack patients are admitted. Identify each case below as a binomial or negative binomial case, and compute the probability.[36]

(a) What is the probability the hospital will admit a heart attack patient on exactly three days this week?

(b) What is the probability the second day with a heart attack patient will be the fourth day of the week?

(c) What is the probability the fifth day of next month will be the first day with a heart attack patient?

[36] In each part, $p = 0.7$. (a) The number of days is fixed, so this is binomial. The parameters are $k = 3$ and $n = 7$: 0.097. (b) The last "success" (admitting a heart attack patient) is fixed to the last day, so we should apply the negative binomial distribution. The parameters are $k = 2$, $n = 4$: 0.132. (c) This problem is negative binomial with $k = 1$ and $n = 5$: 0.006. Note that the negative binomial case when $k = 1$ is the same as using the geometric distribution.

4.5 Poisson distribution

EXAMPLE 4.50

There are about 8 million individuals in New York City. How many individuals might we expect to be hospitalized for acute myocardial infarction (AMI), i.e. a heart attack, each day? According to historical records, the average number is about 4.4 individuals. However, we would also like to know the approximate distribution of counts. What would a histogram of the number of AMI occurrences each day look like if we recorded the daily counts over an entire year?

(E)

A histogram of the number of occurrences of AMI on 365 days for NYC is shown in Figure 4.12.[37] The sample mean (4.38) is similar to the historical average of 4.4. The sample standard deviation is about 2, and the histogram indicates that about 70% of the data fall between 2.4 and 6.4. The distribution's shape is unimodal and skewed to the right.

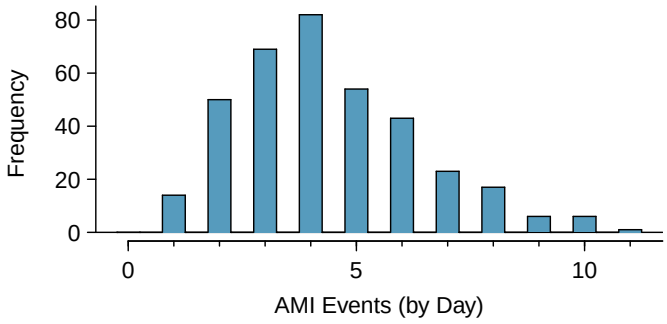


Figure 4.12: A histogram of the number of occurrences of AMI on 365 separate days in NYC.

The **Poisson distribution** is often useful for estimating the number of events in a large population over a unit of time. For instance, consider each of the following events:

- having a heart attack,
- getting married, and
- getting struck by lightning.

The Poisson distribution helps us describe the number of such events that will occur in a day for a fixed population if the individuals within the population are independent. The Poisson distribution could also be used over another unit of time, such as an hour or a week.

The histogram in Figure 4.12 approximates a Poisson distribution with rate equal to 4.4. The **rate** for a Poisson distribution is the average number of occurrences in a mostly-fixed population per unit of time. In Example 4.50, the time unit is a day, the population is all New York City residents, and the historical rate is 4.4. The parameter in the Poisson distribution is the rate – or how many events we expect to observe – and it is typically denoted by λ (the Greek letter *lambda*) or μ. Using the rate, we can describe the probability of observing exactly k events in a single unit of time.

[37] These data are simulated. In practice, we should check for an association between successive days.

POISSON DISTRIBUTION

Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate λ. Then

$$P(\text{observe } k \text{ events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page **??**. The letter $e \approx 2.718$ is the base of the natural logarithm. The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$, respectively.

We will leave a rigorous set of conditions for the Poisson distribution to a later course. However, we offer a few simple guidelines that can be used for an initial evaluation of whether the Poisson model would be appropriate.

A random variable may follow a Poisson distribution if we are looking for the number of events, the population that generates such events is large, and the events occur independently of each other.

Even when events are not really independent – for instance, Saturdays and Sundays are especially popular for weddings – a Poisson model may sometimes still be reasonable if we allow it to have a different rate for different times. In the wedding example, the rate would be modeled as higher on weekends than on weekdays. The idea of modeling rates for a Poisson distribution against a second variable such as `dayOfTheWeek` forms the foundation of some more advanced methods that fall in the realm of **generalized linear models**. In Chapters 8 and 9, we will discuss a foundation of linear models.

Chapter 5

Foundations for inference

Statistical inference is concerned primarily with understanding the uncertainty of parameter estimates. While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. We start with a familiar topic: the notion of using a sample proportion as our estimate of a population proportion. Next, we create what's called a confidence interval, which is a range of values where the true population value is likely to lie. Finally, we introduce a hypothesis testing framework, which allows us to formally evaluate claims about the population, such as *a survey shows a candidate has a majority of support* of the voting population (whether the proportion that supports is greater than 0.5).

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

5.1 Point estimates and sampling variability

Companies such as Pew Research frequently conduct polls as a way to understand the state of public opinion or knowledge on many topics, including politics, scientific understanding, brand recognition, and more. The ultimate goal in taking a poll is generally to use the responses to estimate the opinion or knowledge of the broader population.

5.1.1 Point estimates and error

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a **point estimate** of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the **parameter** of interest, and when the parameter is a proportion, it is often denoted by p, and we often refer to the sample proportion as $\hat{p}$ (pronounced *p-hat*[1]). [Joke in the footnote okay?] Unless we collect responses from every individual in the sample, p remains unknown, and we use $\hat{p}$ as our estimate of p.

The difference we observe from the poll versus the parameter is called the **error** in the estimate. Generally, the error consists of two aspects: sampling error and bias.

Sampling error, sometimes called *sampling uncertainty*, describes how much an estimate will tend to vary from one sample to the next. For instance, the estimate from one sample might be 1% too low while in another it may be 3% too high. Much of statistics, including much of this book, is focused on understanding and quantifying sampling error, and we will find it useful to consider a sample's size to help us quantify this error; the **sample size** is often represented by the letter n.

Bias describes a systematic tendency to over- or under-estimate the true population value. For example, if we were taking a student poll asking about support for a new college stadium, we'd probably get a biased estimate of the stadium's level of student support by wording the question as, *Do you support your school by supporting funding for the new stadium?* We try to minimize bias through thoughtful data collection procedures, which were discussed in Chapter 1 and are the topic of many other books.

5.1.2 Understanding the variability of a point estimate

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest.[2] If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%? We want to understand, *how does the sample proportion $\hat{p}$ behave when the true population proportion is 0.88.*[3] Let's find out! We can simulate responses we would get from a simple random sample of 1000 American adults, which is only possible because we know the actual support expanding solar energy to be 0.88. Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write "support" on 88% of them and "not" on the other 12%.

2. Mix up the pieces of paper and pull out 1000 pieces to represent our sample of 1000 American adults.

3. Compute the fraction of the sample that say "support".

Any volunteers to conduct this simulation? Probably not. Running this simulation with 250 million pieces of paper would be time-consuming and very costly, but we can simulate it using computer code; we've written a short program in Figure 5.1 in case you are curious what the computer code

[1] Not to be confused with *phat*, the slang term used for something cool, like this book.

[2] We haven't actually conducted a census to measure this value perfectly. However, a very large sample has suggested the actual level of support is about 88%.

[3] Note: 88% written as a proportion would be 0.88. It is common to switch between proportion and percent.

looks like. In this simulation, the sample gave a point estimate of $\hat{p}_1 = 0.894$. We know the population proportion for the simulation was $p = 0.88$, so we know the estimate had an error of $0.894 - 0.88 = +0.014$.

```
# 1. Create a set of 250 million entries, where 88% of them are "support"
#    and 12% are "not".
pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size), rep("not", 0.12 * pop_size))

# 2. Sample 1000 entries without replacement.
sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support", then divide by
#    the sample size.
sum(sampled_entries == "support") / 1000
```

Figure 5.1: For those curious, this is code for a single $\hat{p}$ simulation using the statistical software called **R**. Each line that starts with **#** is a **code comment**, which is used to describe in regular language what the code is doing.

One simulation isn't enough to get a great sense of the distribution of estimates we might expect in the simulation, so we should run more simulations. In a second simulation, we get $\hat{p}_2 = 0.885$, which has an error of +0.005. In another, $\hat{p}_3 = 0.878$ for an error of -0.002. And in another, an estimate of $\hat{p}_4 = 0.859$ with an error of -0.021. With the help of a computer, we've run the simulation 10,000 times and created a histogram of the results from all 10,000 simulations in Figure 5.2. This distribution of sample proportions is called a **sampling distribution**. We can characterize this sampling distribution as follows:

Center. The center of the distribution is $\bar{x}_{\hat{p}} = 0.880$, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straightforward sampling strategy that helps avoid sampling bias.

Spread. The standard deviation of the distribution is $s_{\hat{p}} = 0.010$. When we're talking about a sampling distribution or the variability of a point estimate, we typically use the term **standard error** rather than *standard deviation*, and the notation $SE_{\hat{p}}$ is used for the standard error associated with the sample proportion.

Shape. The distribution is symmetric and bell-shaped, and it *resembles a normal distribution.*

These findings are encouraging! When the population proportion is $p = 0.88$ and the sample size is $n = 1000$, the sample proportion $\hat{p}$ tends to give a pretty good estimate of the population proportion. We also have the interesting observation that the histogram resembles a normal distribution.

SAMPLING DISTRIBUTIONS ARE NEVER OBSERVED, BUT WE KEEP THEM IN MIND

In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution. Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

EXAMPLE 5.1

If we used a much smaller sample size of $n = 50$, would you guess that the standard error for $\hat{p}$ would be larger or smaller than when we used $n = 1000$?

(E)

Intuitively, it seems like more data is better than less data, and generally that is correct! The typical error when $p = 0.88$ and $n = 50$ would be larger than the error we would expect when $n = 1000$.

Example 5.1 highlights an important property we will see again and again: a bigger sample tends to provide a more precise point estimate than a smaller sample.

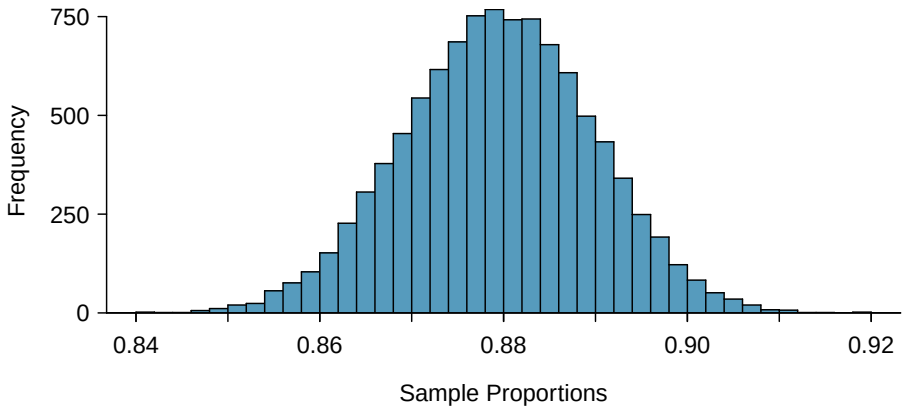


Figure 5.2: A histogram of 10,000 sample proportions, where each sample is taken from a population where the population proportion is 0.88 and the sample size is $n = 1000$.

5.1.3 Central Limit Theorem

The distribution in Figure 5.2 looks an awful lot like a normal distribution. That is no anomaly; it is the result of a general principle called the **Central Limit Theorem**.

CENTRAL LIMIT THEOREM AND THE SUCCESS-FAILURE CONDITION

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:[4]

$$\mu_{\hat{p}} = p \qquad\qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.

The Central Limit Theorem is incredibly important, and it provides a foundation for much of statistics. As we begin applying the Central Limit Theorem, be mindful of the two technical conditions: the observations must be independent, and the sample size must be sufficiently large such that $np \geq 10$ and $n(1-p) \geq 10$.

[4]Technically this formula is the standard deviation of $\hat{p}$, and the standard error is the estimated version using $\hat{p}$. However, to keep terminology simpler – since in every case of inference we do not know the value of p – we use the same name.

EXAMPLE 5.2

Earlier we estimated the mean and standard error of $\hat{p}$ using simulated data when $p = 0.88$ and $n = 1000$. Confirm that the distribution is approximately normal.

Independence. There are $n = 1000$ observations for each sample proportion $\hat{p}$, and each of those observations are independent draws. *The most common way for observations to be considered independent is if they are from a simple random sample.*

E

Success-failure condition. We can confirm the sample size is sufficiently large by checking the success-failure condition and confirming the two calculated values are greater than 10:

$$np = 1000 \times 0.88 = 880 \geq 10 \qquad n(1-p) = 1000 \times (1 - 0.88) = 120 \geq 10$$

Both of the independence and success-failure conditions are satisfied, so the Central Limit Theorem applies and the normal distribution is reasonable in this context.

HOW TO VERIFY SAMPLE OBSERVATIONS ARE INDEPENDENT

Subjects in an experiment are considered independent if they undergo random assignment to the treatment groups.

If the observations are from a simple random sample, then they are independent.

If a sample is from a seemingly random process, e.g. an occasional error on an assembly line, checking independence is more difficult. In this case, use your best judgement.

An additional condition that is sometimes added for samples from a population is that they are no larger than 10% of the population. When the sample exceeds 10% of the population size, the methods we discuss tend to overestimate the sampling error slightly versus what we would get using more advanced methods.[5] This is very rarely an issue, and when it is an issue, our methods tend to be conservative, so we consider this additional check as optional.

EXAMPLE 5.3

Compute the theoretical mean and standard error of $\hat{p}$ when $p = 0.88$ and $n = 1000$, according to the Central Limit Theorem.

The mean of the $\hat{p}$'s is simply the population proportion: $\mu_{\hat{p}} = 0.88$.

E

The calculation of the standard error of $\hat{p}$ uses the following formula:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88(1-0.88)}{1000}} = 0.010$$

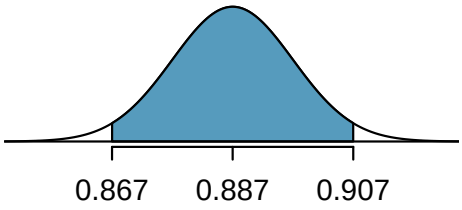
[5] For example, we could use what's called the **finite population correction factor**: if the sample is of size n and the population size is N, then we can multiple the typical standard error formula by $\sqrt{\frac{N-n}{N-1}}$ to obtain a smaller, more precise estimate of the actual standard error. When $n < 0.1 \times N$, this correction factor is relatively small.

EXAMPLE 5.4

Estimate how frequently the sample proportion $\hat{p}$ should be within 0.02 (2%) of the population value, $p = 0.88$. Based on Examples 5.2 and 5.3, we know that the distribution is approximately $N(\mu_{\hat{p}} = 0.88, SE_{\hat{p}} = 0.010)$.

After so much practice in Section 4.1, this normal distribution example will hopefully feel familiar! We would like to understand the fraction of $\hat{p}$'s between 0.86 and 0.90:

E



With $\mu_{\hat{p}} = 0.887$ and $SE_{\hat{p}} = 0.010$, we can compute the Z-score for both the left and right cutoffs:

$$Z_{0.86} = \frac{0.86 - 0.88}{0.010} = -2 \qquad Z_{0.90} = \frac{0.90 - 0.88}{0.010} = 2$$

We can use either statistical software, a graphing calculator, or a table to find the areas to the tails, and in any case we will find that they are each 0.0228. The total tail areas are $2 \times 0.0228 = 0.0456$, which leaves the shaded area of 0.9544. That is, about 95.44% of the sampling distribution in Figure 5.2 is within ± 0.02 of the population proportion, $p = 0.88$.

GUIDED PRACTICE 5.5

G In Example 5.1 we discussed how a smaller sample would tend to produce a less reliable estimate. Explain how this intuition is reflected in the formula for $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.[6]

5.1.4 Applying the Central Limit Theorem to a real-world setting

We do not actually know the population proportion unless we conduct an expensive poll of all individuals in the population. Our earlier value of $p = 0.88$ was based on a Pew Research conducted a poll of 1000 American adults that found $\hat{p} = 0.887$ of them favored expanding solar energy. The researchers might have wondered: does the sample proportion from the poll approximately follow a normal distribution? We can the conditions from the Central Limit Theorem:

Independence. The poll is a simple random sample of American adults, which means that the observations are independent.

Success-failure condition. To check this condition, we need the population proportion, p, to check if both np and $n(1-p)$ are greater than 10. However, we do not actually know p, which is exactly why the pollsters would take a sample! In cases like these, we often use $\hat{p}$ as our next best way to check the success-failure condition:

$$n\hat{p} = 1000 \times 0.887 = 887 \qquad n(1-\hat{p}) = 1000 \times (1 - 0.887) = 113$$

The sample proportion $\hat{p}$ acts as a reasonable substitute for p during this check, and each value in this case is well above the minimum of 10.

This **substitution approximation** of using $\hat{p}$ in place of p is also useful when computing the

[6] Since the sample size n is in the denominator (on the bottom) of the fraction, a bigger sample size means the entire expression when calculated will tend to be smaller. That is, a larger sample size would correspond to a smaller standard error.

standard error of the sample proportion:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.887(1-0.887)}{1000}} = 0.010$$

This substitution technique is sometimes referred to as the "plug-in principle". In this case, $SE_{\hat{p}}$ didn't change enough to be detected using only 3 decimal places versus when we completed the calculation with 0.88 earlier. The computed standard error tends to be reasonably stable even when observing slightly different proportions in one sample or another.

5.1.5 More details regarding the Central Limit Theorem

We've applied the Central Limit Theorem in numerous examples so far this chapter:

> *When observations are independent and the sample size is sufficiently large, the distribution of $\hat{p}$ resembles a normal distribution with*
>
> $$\mu_{\hat{p}} = p \qquad\qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$
>
> *The sample size is considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$.*

In this section, we'll explore the success-failure condition and seek to better understand the Central Limit Theorem.

An interesting question to answer is, *what happens when $np < 10$ or $n(1-p) < 10$?* As we did in Section 5.1.2, we can simulate drawing samples of different sizes where, say, the true proportion is $p = 0.25$. Here's a sample of size 10:

no, no, yes, yes, no, no, no, no, no, no

In this sample, we observe a sample proportion of yeses of $\hat{p} = \frac{2}{10} = 0.2$. We can simulate many such proportions to understand the sampling distribution of $\hat{p}$ when $n = 10$ and $p = 0.25$, which we've plotted in Figure 5.3 alongside a normal distribution with the same mean and variability. These distributions have a number of important differences.

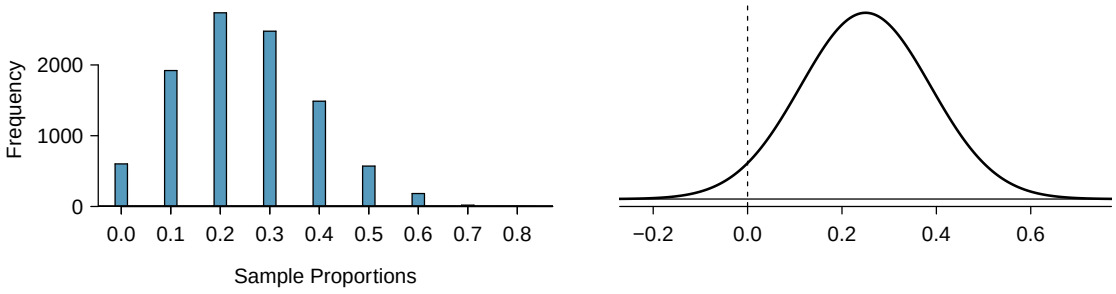


Figure 5.3: Left: simulations of $\hat{p}$ when the sample size is $n = 10$ and the population proportion is $p = 0.25$. Right: a normal distribution with the same mean (0.25) and standard deviation (0.137).

	Unimodal?	Smooth?	Symmetric?
Normal: $N(0.25, 0.14)$	**Yes**	**Yes**	**Yes**
$n = 10$, $p = 0.25$	**Yes**	*No*	*No*

Notice that the success-failure condition was not satisfied when $n = 10$ and $p = 0.25$:

$$np = 10 \times 0.25 = 2.5 \qquad\qquad n(1-p) = 10 \times 0.75 = 7.5$$

This single sampling distribution does not show that the success-failure condition is the perfect guideline, but we have found that the guideline did correctly identify that the normal distribution might not be appropriate.

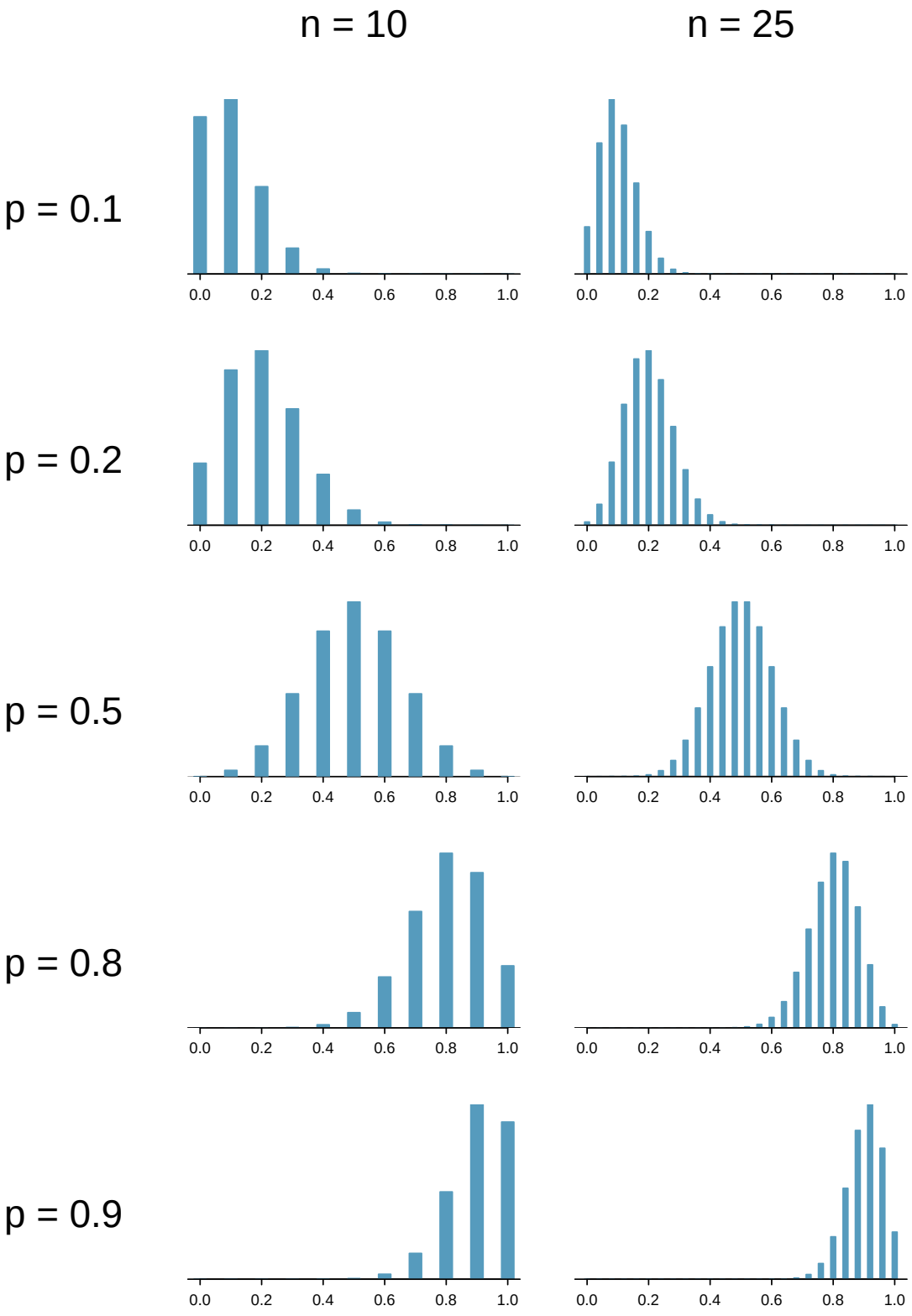


Figure 5.4: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 10$ and $n = 25$.

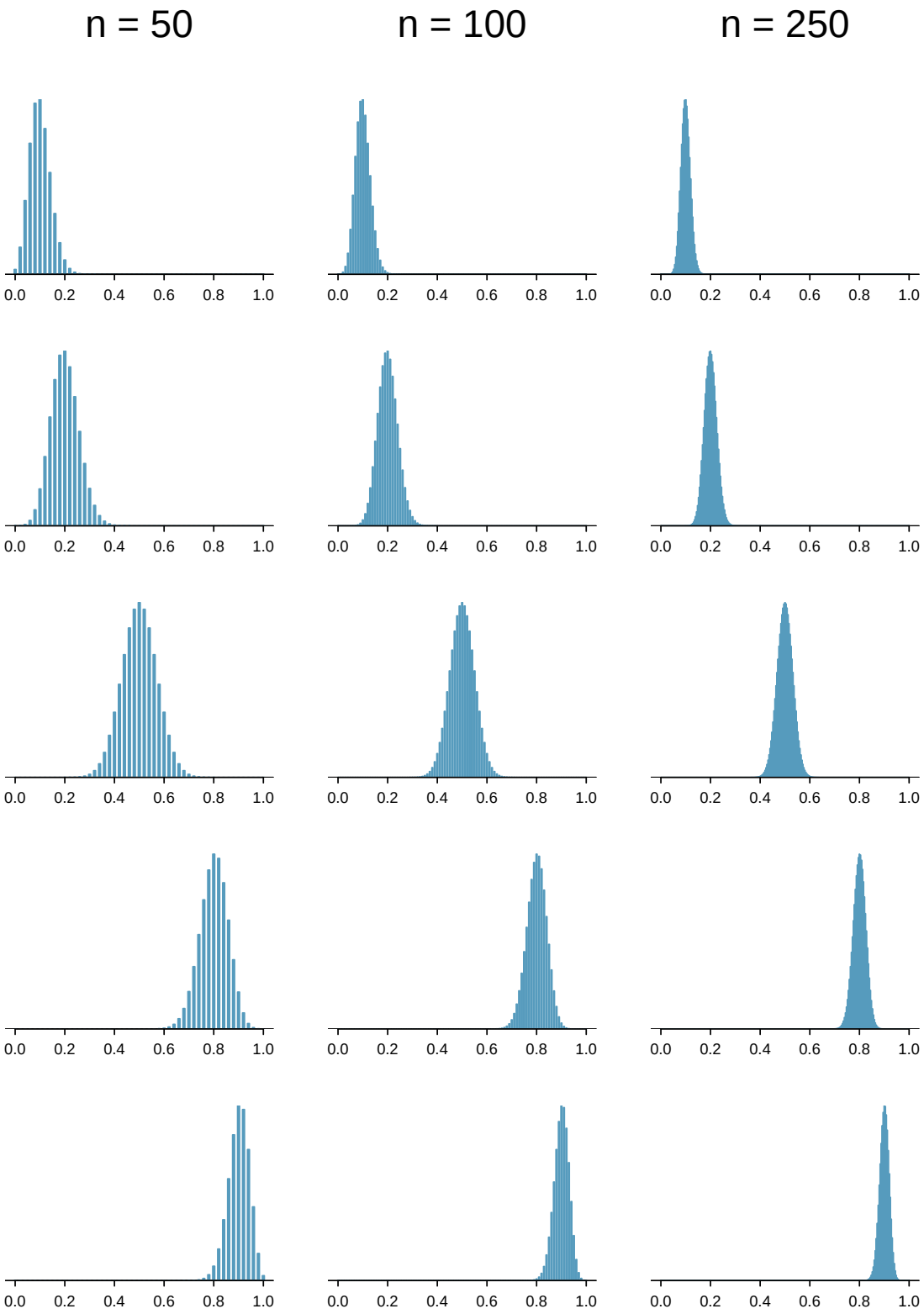


Figure 5.5: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 50$, $n = 100$, and $n = 250$.

We can complete several additional simulations, shown in Figures 5.4 and 5.5. We can see some trends:

1. When either np or $n(1-p)$ is small, the distribution is more **discrete**, i.e. *not continuous*.
2. When np or $n(1-p)$ is smaller than 10, the skew in the distribution is more noteworthy.
3. The larger both np *and* $n(1-p)$, the more normal the distribution. This may be a little harder to see for the larger sample size in these plots as the variability also becomes much smaller.
4. When np and $n(1-p)$ are both very large, the distribution's discreteness is hardly evident, and the distribution looks much more like a normal distribution.

So far we've only focused on the skew and discreteness of the distributions. We haven't considered how the mean and standard error of the distributions change. Take a moment to look back at the graphs, and pay attention to three things:

1. The centers of the distribution are always at the population proportion, p, that was used to generate the simulation. Because the sampling distribution of $\hat{p}$ is always centered at the population parameter p, it means the sample proportion $\hat{p}$ is **unbiased** when the data are independent and drawn from such a population.
2. For a particular population proportion p, the variability in the sampling distribution decreases as the sample size n becomes larger. This will likely align with your intuition: an estimate based on a larger sample size will tend to be more accurate.
3. For a particular sample size, the variability will be largest when $p = 0.5$. The differences may be a little subtle, so take a close look. This reflects the role of the proportion p in the standard error formula: $SE = \sqrt{\frac{p(1-p)}{n}}$. The standard error is largest when $p = 0.5$.

At no point will the distribution of $\hat{p}$ look *perfectly* normal, since $\hat{p}$ will always be take discrete values (x/n). It is always a matter of degree, and we will use the standard success-failure condition with minimums of 10 for np and $n(1-p)$ as our guideline within this book.

5.2 Confidence intervals for a sample proportion

The sample proportion $\hat{p}$ provides a single plausible value for the population proportion p. However, the sample proportion isn't perfect and will have some *standard error* associated with it. Instead of supplying just this point estimate of the population proportion, a next logical step would be to provide a plausible *range of values* for the population proportion.

5.2.1 Capturing the population parameter

Using only a point estimate is like fishing in a murky lake with a spear. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish. A **confidence interval** is like fishing with a net, and it represents a range of plausible values where we are likely to find the population parameter.

If we report a point estimate $\hat{p}$, we probably will not hit the exact population proportion. On the other hand, if we report a range of plausible values, representing a confidence interval, we have a good shot at capturing the parameter.

GUIDED PRACTICE 5.6

(G) If we want to be very certain we capture the population proportion in an interval, should we use a wider interval or a smaller interval?[7]

5.2.2 An approximate 95% confidence interval

Our sample proportion $\hat{p}$ is the most plausible value of the population proportion, so it makes sense to build a confidence interval around this point estimate. The standard error provides a guide for how large we should make the confidence interval.

The standard error represents the standard deviation of the point estimate, and when the Central Limit Theorem conditions are satisfied, we also know that the point estimate closely follows a normal distribution. In a normal distribution, about 95% of the data is within 2 standard deviations of the mean. Using this principle, we can construct a confidence interval that extends 2 standard errors from the sample proportion to be **95% confident** that the interval captures the population proportion:

$$\begin{aligned} \text{point estimate} &\pm 2 \times SE \\ \hat{p} &\pm 2 \times SE_{\hat{p}} \end{aligned}$$

But what does "95% confident" mean? Suppose we took many samples and built a 95% confidence interval from each. Then about 95% of those intervals would contain the parameter, p. Figure 5.6 shows the process of creating 25 intervals from 25 samples from the simulation in Section 5.1.2, where 24 of the resulting confidence intervals contain the simulation's population proportion of $p = 0.88$, and one interval does not.

EXAMPLE 5.7

In Figure 5.6, one interval does not contain $p = 0.88$. Does this imply that the population proportion used in the simulation could not have been $p = 0.88$?

(E) Just as some observations naturally occur more than 2 standard deviations from the mean, some point estimates will be more than 2 standard errors from the parameter of interest. A confidence interval only provides a plausible range of values. While we might say other values are implausible based on the data, this does not mean they are impossible.

[7] 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.

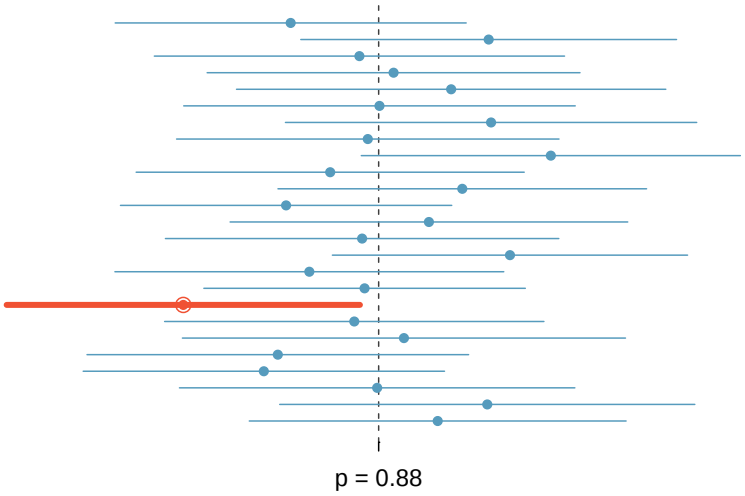


Figure 5.6: Twenty-five point estimates and confidence intervals from the simulations in Section 5.1.2. These intervals are shown relative to the population proportion $p = 0.88$. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been bolded.

While about 95% of the data is within 2 standard deviations in a normal distribution, it would be more precise to use a value of 1.96 standard deviations as the area under the normal curve within 1.96 standard deviations is closer to 0.95. This more precise value is what is used to construct a 95% confidence interval.

95% CONFIDENCE INTERVAL FOR A PARAMETER

When the distribution of a point estimate qualifies for the Central Limit Theorem and therefore closely follows a normal distribution, we can construct a 95% confidence interval as

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

EXAMPLE 5.8

In Section 5.1 we learned about a Pew Research poll where 88.7% of a random sample of 1000 American adults supported expanding the role of solar power. Compute and interpret a 95% confidence interval for the population proportion.

(E)

We earlier confirmed that $\hat{p}$ follows a normal distribution and has a standard error of $SE_{\hat{p}} = 0.010$. To compute the 95% confidence interval, plug the point estimate $\hat{p} = 0.887$ and standard error into the 95% confidence interval formula:

$$\hat{p} \pm 1.96 \times SE_{\hat{p}} \quad \rightarrow \quad 0.887 \pm 1.96 \times 0.010 \quad \rightarrow \quad (0.8674, 0.9066)$$

We are 95% confident that the actual proportion of American adults who support expanding solar power is between 86.7% and 90.7%. (It's common to round to the nearest percentage point or nearest tenth of a percentage point when reporting a confidence interval.)

5.2.3 Changing the confidence level

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

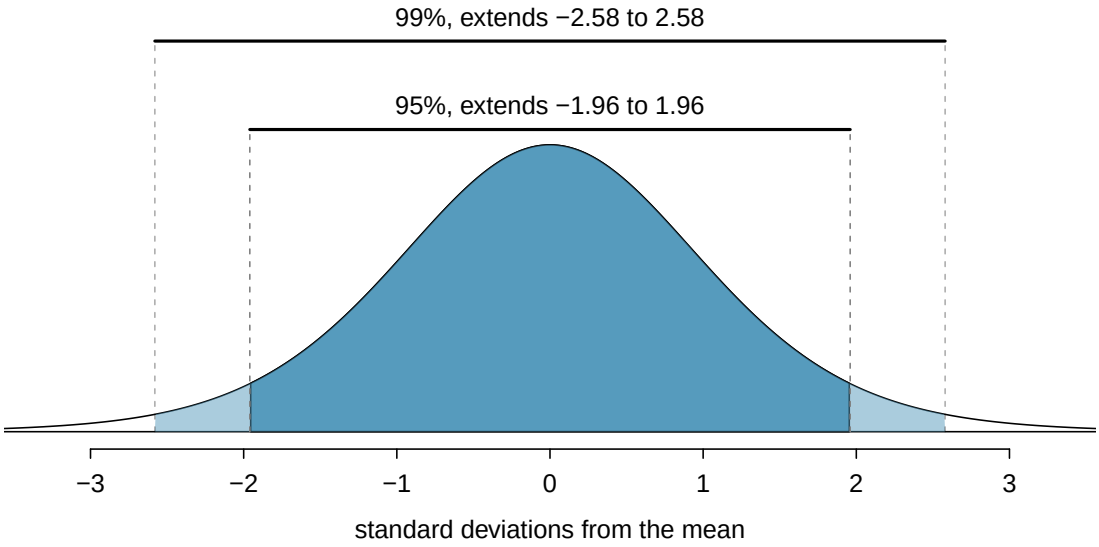


Figure 5.7: The area between -$z^\star$ and $z^\star$ increases as $z^\star$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

The 95% confidence interval structure provides guidance in how to make intervals with different confidence levels. The general 95% confidence interval for a point estimate that follows the normal distribution is

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

There are three components to this interval: the point estimate, "1.96", and the standard error. The choice of $1.96 \times SE$ was based on capturing 95% of the data since the estimate is within 1.96 standard errors of the parameter about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

GUIDED PRACTICE 5.9

G If X is a normally distributed random variable, what is the probability of the value X being within 2.58 standard deviations of the mean?[8]

Guided Practice 5.9 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean. To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58. That is, the formula for a 99% confidence interval is

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

This approach – using the Z-scores in the normal model to compute confidence levels – is appropriate when a point estimate such as $\hat{p}$ is associated with a normal distribution. For some other point estimates, the normal model is not a good fit; in these cases, we'll use alternative distributions that better represent the sampling distribution.

[8]This is equivalent to asking how often the Z-score will be larger than -2.58 but less than 2.58. For a picture, see Figure 5.7. To determine this probability, we can use statistical software, a calculator, or a table to look up -2.58 and 2.58 for the normal distribution: 0.0049 and 0.9951. Thus, there is a $0.9951 - 0.0049 \approx 0.99$ probability that an unobserved normal random variable X will be within 2.58 standard deviations of μ.

CONFIDENCE INTERVAL USING ANY CONFIDENCE LEVEL

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

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

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

Figure 5.7 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 distribution corresponds to the confidence level.

MARGIN OF ERROR

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

EXAMPLE 5.10

Use the data in Example 5.8 to create a 90% confidence interval for the proportion of American adults that support expanding the use of solar power. We have already verified conditions for normality.

E

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 do this using a graphing calculator, statistical software, or a probability table by looking for an upper tail of 5% (the other 5% is in the lower tail): $z^\star = 1.65$. The 90% confidence interval can then be computed as

$$\hat{p} \pm 1.65 \times SE_{\hat{p}} \quad \rightarrow \quad 0.887 \pm 1.65 \times 0.0100 \quad \rightarrow \quad (0.8705, 0.9035)$$

That is, we are 90% confident that 87.1% to 90.4% of American adults supported the expansion of solar power in 2018.

CONFIDENCE INTERVAL FOR A SINGLE PROPORTION

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\hat{p}$ and n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one-proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure condition.

Calculate. If the conditions hold, compute SE using $\hat{p}$, find $z^\star$, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

5.2.4 More case studies

In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014.

EXAMPLE 5.11

What is the point estimate in this case, and is it reasonable to use the normal distribution to model that point estimate?

(E)

The point estimate, based on a sample of size $n = 1042$, is $\hat{p} = 0.82$. To check whether $\hat{p}$ can be reasonably modeled using the normal distribution, we check independence (the poll is based on a simple random sample) and the success-failure condition ($1042 \times \hat{p} \approx 854$ and $1042 \times (1 - \hat{p}) \approx 188$, both easily greater than 10). With the conditions met, we are assured that the sampling distribution of $\hat{p}$ can be reasonably modeled using a normal distribution.

EXAMPLE 5.12

Estimate the standard error of $\hat{p} = 0.82$ from the Ebola survey.

(E)

We'll use the substitution approximation of $p \approx \hat{p} = 0.82$ to compute the standard error:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.82(1-0.82)}{1042}} = 0.012$$

EXAMPLE 5.13

Construct a 95% confidence interval for p, the proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.

(E)

Using the standard error $SE = 0.012$ from Example 5.12, the point estimate 0.82, and $z^\star = 1.96$ for a 95% confidence level, the confidence interval is

$$\text{point estimate} \ \pm \ z^\star \times SE \quad \rightarrow \quad 0.82 \ \pm \ 1.96 \times 0.012 \quad \rightarrow \quad (0.796, 0.844)$$

We are 95% confident that the proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patient was between 0.796 and 0.844.

GUIDED PRACTICE 5.14

(G)

Answer the following two questions about the confidence interval from Example 5.13:[9]

(a) What does 95% confident mean in this context?

(b) Do you think the confidence interval is still valid for the opinions of New Yorkers today?

GUIDED PRACTICE 5.15

In the poll by Pew Research asking about solar energy, the researchers also inquired about other forms of energy. Support for expanding wind turbines received support from 84.8% of the 1000 respondents.[10]

(G)

(a) Is the normal approximation reasonable for the proportion of US adults who support expanding wind turbines?

(b) Create a 99% confidence interval for the level of American support for expanding the use of wind turbines for power generation.

[9] (a) If we took many such samples and computed a 95% confidence interval for each, then about 95% of those intervals would contain the actual proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.
(b) Not necessarily. The poll was taken at a time where there was a huge public safety concern. Now that people have had some time to step back, they may have changed their opinions. We would need to run a new poll if we wanted to get an estimate of the current proportion of New York adults who would support such a quarantine period.

5.2.5 Interpreting confidence intervals

In each of the examples, we described the confidence intervals by putting them into the context of the data and also using somewhat formal language:

Solar. We are 90% confident that 87.1% to 90.4% of American adults support the expansion of solar power in 2018.

Ebola. We are 95% confident that the proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patient was between 0.796 and 0.844.

Wind Turbine. We are 99% confident the proportion of Americans adults that support expanding the use of wind turbines is between 81.9% and 87.7% in 2018.

First, notice that the statements are always about the population parameter, which considers *all* American adults for the energy polls or *all* New York adults for the quarantine poll.

We also avoided another common mistake: *incorrect* language might try to describe the confidence interval as capturing the population parameter with a certain probability. Making a probability interpretation is a common error: 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 important consideration of confidence intervals is that they are *only about the population parameter*. A confidence interval says nothing about individual observations, a proportion of the observations, or point estimates. Confidence intervals only provide a plausible range for population parameters.

Lastly, keep in mind the methods we discussed only apply to sampling error, not to bias. If a data set is collected in a way that will tend to result in a bias that tends to systematically underestimate (or over-estimate) the population parameter, the techniques we have discussed will not address that problem. Instead, we rely on careful data collection procedures to help protect against bias in the examples we have considered, which is a common practice employed by data scientists to combat bias.

[10](a) We check independence, which is okay since this survey was a simple random sample, and also the success-failure condition ($1000 \times 0.848 = 848$ and $1000 \times 0.152 = 152$ are both at least 10). Since both conditions are satisfied, $\hat{p} = 0.848$ can be modeled using the normal distribution.
(b) Guided Practice 5.15 confirmed that that $\hat{p}$ closely follows a normal distribution, so we can use the C.I. formula:

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

In this case, the point estimate is $\hat{p} = 0.848$. For a 99% confidence interval, $z^\star = 2.58$. Computing the standard error: $SE_{\hat{p}} = \sqrt{\frac{0.848(1-0.848)}{1000}} = 0.0114$. Finally, we compute the interval as $0.848 \pm 2.58 \times 0.0114 \rightarrow (0.8186, 0.8774)$. It is also important to *always* provide an interpretation for the interval: we are 99% confident the proportion of American adults that support expanding the use of wind turbines in 2018 is between 81.9% and 87.7%.

5.3 Hypothesis testing for a proportion

[Update OpenIntro's site to handle all new redirect links, many of which are currently broken in this edition.]

The following question comes from a book written by Hans Rosling, Anna Rosling Rönnlund, and Ola Rosling called *Factfulness*:

> *How many of the world's 1 year old children today have been vaccinated against some disease:*
>
> a. *20%*
> b. *50%*
> c. *80%*

Write down what your answer (or guess), and when you're ready, find the answer in the footnote.[11]

In this section, we'll be exploring how people with a 4-year college degree perform on this and other world health questions.

5.3.1 Hypothesis testing framework

We're interested in understanding whether people know much about world health and development. If we take a multiple choice world health question, then we might like to understand if

H_0: People never learn these particular topics and their responses are simply equivalent to random guesses.

H_A: Folks have knowledge that helps them do better than random guessing, or perhaps, they have false knowledge that leads them to actually do worse than random guessing.

These competing ideas are called **hypotheses**. We call H_0 the null hypothesis and H_A the alternative hypothesis. When there is a subscript 0 like in H_0, data scientists pronounce it as "nought" (e.g. H_0 is pronounced "H-nought").

NULL AND ALTERNATIVE HYPOTHESES

The **null hypothesis (H_0)** often represents 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.

Our job as data scientists is to play the role of a skeptic: before we buy into the alternative hypothesis, we need to see strong supporting evidence.

The null hypothesis often represents a skeptical position or a perspective of "no difference". In our first example, we'll consider whether the typical person does any different than random guessing on Roslings' question about infant vaccinations.

The alternative hypothesis generally represents a new or stronger perspective. In the case of the question about infant vaccinations, it would certainly be interesting to learn whether people do better than random guessing, since that would mean that the typical person knows something about world health statistics. It would also be very interesting if we learned that people do *worse* than random guessing, which would suggest people believe incorrect information about world health.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative. The hallmarks of hypothesis testing are also found in the US court system.

[11] The correct answer is (c): 80% of the world's 1 year olds have been vaccinated against some disease.

GUIDED PRACTICE 5.16

A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?[12]

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. 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.

When considering Roslings' question about infant vaccination, the null hypothesis represents the notion that the people we will be considering – college-educated adults – are as accurate as random guessing. That is, the proportion p of respondents who pick the correct answer, that 80% of 1 year olds have been vaccinated against some disease, is about 33.3% (or 1-in-3 if wanting to be perfectly precise). The alternative hypothesis is that this proportion is something other than 33.3%. While it's helpful to write these hypotheses in words, it can be useful to write them using mathematical notation:

H_0**:** $p = 0.333$

H_A**:** $p \neq 0.333$

In this hypothesis setup, we want to make a conclusion about the population parameter p. The value we are comparing the parameter to is called the **null value**, which in this case is 0.333. It's common to label the null value with the same symbol as the parameter but with a subscript '0'. That is, in this case, the null value is $p_0 = 0.333$ (pronounced "p-nought equals 0.333").

EXAMPLE 5.17

It may seem impossible that the proportion of people who get the correct answer is *exactly* 33.3%. If we don't believe the null hypothesis, should we simply reject it?

No. While we may not buy into the notion that the proportion is exactly 33.3%, the hypothesis testing framework requires that there be strong evidence before we reject the null hypothesis and conclude something more interesting.

After all, even if we don't believe the proportion is *exactly* 33.3%, that doesn't really tell us anything useful! We would still be stuck with the original question: do people do better or worse than random guessing on Roslings' question? Without data that strongly points in one direction or the other, it is both uninteresting and pointless to reject H_0.

GUIDED PRACTICE 5.18

Another example of a real-world hypothesis testing situation is evaluating whether a new drug is better or worse than an existing drug at treating a particular disease. What should we use for the null and alternative hypotheses in this case?[13]

5.3.2 Testing hypotheses using confidence intervals

[Build the `rosling_responses` data set and put it in the R package.]

We will use the `rosling_responses` data set to evaluate the hypothesis test evaluating whether college-educated adults who get the question about infant vaccination correct is different from 33.3%.

[12] The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

[13] The null hypothesis (H_0) in this case is the declaration of *no difference*: the drugs are equally effective. The alternative hypothesis (H_A) is that the new drug performs differently than the original, i.e. it could perform better or worse.

This data set summarizes the answers of 50 college-educated adults. Of these 50 adults, 24% of respondents got the question correct that 80% of 1 year olds have been vaccinated against some disease.

Up until now, our discussion has been philosophical. However, now that we have data, we might ask ourselves: does the data provide strong evidence that the proportion of college-educated adults is different than 33.3%?

We learned in Section 5.1 that there is fluctuation from one sample to another, and it is unlikely that our sample proportion, $\hat{p}$, will exactly equal p, but we want to make a conclusion about p. We have a nagging concern: is this deviation of 24% from 33.3% simply due to chance, or does the data provide strong evidence that the population proportion is different from 33.3%?

In Section 5.2, we learned how to quantify the uncertainty in our estimate using confidence intervals. The same method for measuring variability can be useful for the hypothesis test.

EXAMPLE 5.19

Check whether it is reasonable to construct a confidence interval for p using the sample data, and if so, construct a 95% confidence interval.

The conditions are met for $\hat{p}$ to be approximately normal: the data come from a simple random sample (satisfies independence), and $n\hat{p} = 12$ and $n(1-\hat{p}) = 38$ are both at least 10 (success-failure condition).

(E) To construct the confidence interval, we will need to identify the point estimate ($\hat{p} = 0.24$), the critical value for the 95% confidence level ($z^\star = 1.96$), and the standard error of $\hat{p}$ ($SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.060$). With those pieces, the confidence interval for p can be constructed:

$$\hat{p} \pm z^\star \times SE_{\hat{p}}$$
$$0.24 \pm 1.96 \times 0.060$$
$$(0.122, 0.358)$$

We are 95% confident that the proportion of all college-educated adults to correctly answer this particular question about infant vaccination is between 12.2% and 35.8%.

Because the null value in the hypothesis test is $p_0 = 0.333$, which falls within the range of plausible values from the confidence interval, we cannot say the null value is implausible.[14] That is, the data do not provide sufficient evidence to reject the notion that the performance of college-educated adults was different than random guessing, and we do not reject the null hypothesis, H_0.

EXAMPLE 5.20

Explain why we cannot conclude that college-educated adults simply guessed on the infant vaccination question.

(E)

While we failed to reject H_0, that does not necessarily mean the null hypothesis is true. Perhaps there was an actual difference, but we were not able to detect it with the relatively small sample of 50.

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.

[14] Arguably this method is slightly imprecise. As we'll see in a few pages, the standard error is often computed slightly differently in the context of a hypothesis test for a proportion.

GUIDED PRACTICE 5.21

Let's move onto a second question posed by the Roslings:

> *There are 2 billion children in the world today aged 0-15 years old, how many children will there be in year 2100 according to the United Nations?*
>
> *a. 4 billion.*
> *b. 3 billion.*
> *c. 2 billion.*

Set up appropriate hypotheses to evaluate whether college-educated adults are better than random guessing on this question. Also, see if you can guess the correct answer before checking the answer in the footnote![15]

GUIDED PRACTICE 5.22

This time we took a larger sample of 228 college-educated adults, 34 (14.9%) selected the correct answer to the question in Guided Practice 5.21: 2 billion. Can we apply the normal model to the sample proportion to construct a confidence interval in this case?[16]

EXAMPLE 5.23

Compute a 95% confidence interval for the fraction of college-educated adults who answered the children-in-2100 question correctly, and evaluate the hypotheses in Guided Practice 5.21.

To compute the standard error, we'll again use $\hat{p}$ in place of p for the calculation:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.149(1-0.149)}{228}} = 0.024$$

You showed in Guided Practice 5.22 that the normal model may be applied to $\hat{p}$, which ensures a 95% confidence interval may be accurately constructed as

$$\hat{p} \pm z^\star \times SE \quad \rightarrow \quad 0.149 \pm 1.96 \times 0.024 \quad \rightarrow \quad (0.103, 0.195)$$

Because the null value, $p_0 = 0.333$, is not in the confidence interval, a population proportion of 0.333 is implausible and we reject the null hypothesis. That is, the data provide statistically significant evidence that the actual proportion of college adults who get the children-in-2100 question correct is different from random guessing. Because the entire 95% confidence interval is below 0.333, we can conclude college-educated adults do *worse* than random guessing on this question.

One subtle consideration is that we used a 95% confidence interval. What if we had used a 99% confidence level? Or even a 99.9% confidence level? It's possible to come to a different conclusion if using a different confidence level. Therefore, when we make a conclusion based on confidence interval, we should also be sure it is clear what confidence level we used.

The worse-than-random performance on this last question is not a fluke: there are many such world health questions where people do worse than random guessing. In general, the answers suggest that people tend to be more pessimistic about progress than reality suggests. This topic is discussed in much greater detail in the Roslings' book, *Factfulness*.

---

[15] The appropriate hypotheses are:
H_0: the proportion who get the answer correct is the same as random guessing: 1-in-3, or $p = 0.333$.
H_A: the proportion who get the answer correct is different than random guessing, $p \neq 0.333$.
The correct answer to the question is 2 billion. While the world population is projected to increase, the average age is also expected to rise. That is, the majority of the population growth will happen in older age groups, meaning people are projected to live longer in the future across much of the world.

[16] We check both conditions, which are satisfied, so it is reasonable to use the normal distribution for $\hat{p}$:
Independence. Since the data are from a simple random sample, the observations are independent.
Success-failure. We'll use $\hat{p}$ in place of p to check: $n\hat{p} = 34$ and $n(1-\hat{p}) = 194$. Both are greater than 10, so the success-failure condition is satisfied.

5.3.3 Decision errors

Hypothesis tests are not flawless: we can make an incorrect decision in a statistical hypothesis test based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. One key distinction with statistical hypothesis tests is that we have the tools necessary to probabilistically quantify how often we make errors in our conclusions.

Recall that 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, which are summarized in Figure 5.8.

		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

Figure 5.8: 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 5.24

G 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? Figure 5.8 may be useful.[17]

EXAMPLE 5.25

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

E 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 5.26

G How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[18]

Exercises 5.24-5.26 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. That is, if the null hypothesis is true, the significance level indicates how often the data lead us to incorrectly reject H_0. We often write the significance level using α (the Greek letter *alpha*): $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 5.3.5.

[17] If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. Note that a Type 1 Error is only possible if we've rejected the null hypothesis.

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). Note that a Type 2 Error is only possible if we have failed to reject the null hypothesis.

[18] 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.

If we use a 95% confidence interval to evaluate a hypothesis test and the null hypothesis happens to be 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 $\alpha = 0.01$.

A confidence interval is very helpful in determining whether or not to reject the null hypothesis. However, the confidence interval approach isn't always sustainable. In several sections, we will encounter situations where a confidence interval cannot be constructed. For example, if we wanted to evaluate the hypothesis that several proportions are equal, it isn't clear how to construct and compare many confidence intervals altogether.

Next we will introduce a statistic called the *p-value* to help us expand our statistical toolkit, which will enable us to both better understand the strength of evidence and work in more complex data scenarios in later sections.

5.3.4 Formal testing using p-values

[Jon suggested we use a simulation to introduce the p-value. Seems like it could ease some of the concepts.]

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative hypothesis. Statistical hypothesis testing typically uses the p-value method rather than confidence intervals.

P-VALUE

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, in this section the sample proportion, to help compute the p-value and evaluate the hypotheses.

EXAMPLE 5.27

Pew Research asked a random sample of 1000 American adults whether they supported the increased usage of coal to produce energy. Set up hypotheses to evaluate whether a majority of American adults support or oppose the increased usage of coal.

E

The uninteresting result is that there is no majority either way: half of Americans support and the other half oppose expanding the use of coal to produce energy. The alternative hypothesis would be that there is a majority support or oppose (though we do not known which one!) expanding the use of coal. If p represents the proportion supporting, then we can write the hypotheses as

H_0**:** $p = 0.5$

H_A**:** $p \neq 0.5$

In this case, the null value is $p_0 = 0.5$.

When evaluating hypotheses for proportions using the p-value method, we will slightly modify how we check the success-failure condition and compute the standard error for the single proportion case. These changes aren't dramatic, but pay close attention to how we use the null value, p_0.

EXAMPLE 5.28

Pew Research's sample found that 37% of American adults support increased usage of coal. We now wonder, does 37% represent a real difference from the null hypothesis of 50%? What would the sampling distribution of $\hat{p}$ look like if the null hypothesis were true?

If the null hypothesis were true, the population proportion would be the null value, 0.5. We previously learned that the sampling distribution of $\hat{p}$ will be normal when two conditions are met:

Independence. The poll was based on a simple random sample, so independence is satisfied.

Success-failure. Based on the poll's sample size of $n = 1000$, the success-failure condition is met, since

$$np \overset{H_0}{=} 1000 \times 0.5 = 500 \qquad n(1-p) \overset{H_0}{=} 1000 \times (1-0.5) = 500$$

(E) are both at least 10. Note that the success-failure condition was checked using the null value, $p_0 = 0.5$; this is the first procedural difference from confidence intervals.

If the null hypothesis were true, the sampling distribution indicates that a sample proportion based on $n = 1000$ observations would be normally distributed. Next, we can compute the standard error, where we will again use the null value $p_0 = 0.5$ in the calculation:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \overset{H_0}{=} \sqrt{\frac{0.5 \times (1-0.5)}{1000}} = 0.016$$

This marks the other procedural difference from confidence intervals: since the sampling distribution is determined under the null proportion, the null value p_0 was used for the proportion in the calculation rather than $\hat{p}$.

Ultimately, if the null hypothesis were true, then the sample proportion should follow a normal distribution with mean 0.5 and a standard error of 0.016. This distribution is shown in Figure 5.9.

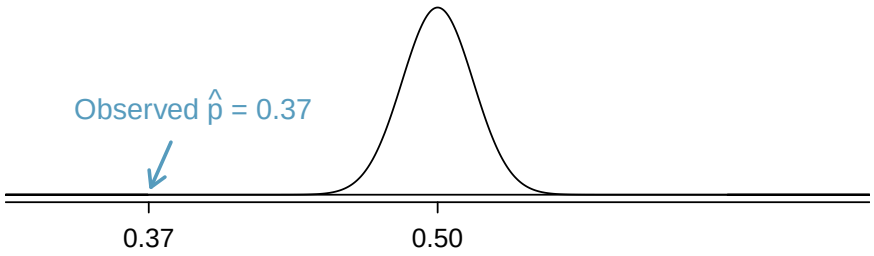


Figure 5.9: If the null hypothesis were true, this normal distribution describes the distribution of $\hat{p}$.

CHECKING SUCCESS-FAILURE AND COMPUTING $SE_{\hat{p}}$ FOR A HYPOTHESIS TEST

When using the p-value method to evaluate a hypothesis test, we check the conditions for $\hat{p}$ and construct the standard error using the null value, p_0, instead of using the sample proportion.

In a hypothesis test with a p-value, we are supposing the null hypothesis is true, which is a different mindset than when we compute a confidence interval. This is why we use p_0 instead of $\hat{p}$ when we check conditions and compute the standard error in this context.

When we identify the sampling distribution under the null hypothesis, it has a special name: the **null distribution**. The p-value represents the probability of the observed $\hat{p}$, or a $\hat{p}$ that is more extreme, if the null hypothesis were true. To find the p-value, we generally find the null distribution, and then we find a tail area in that distribution corresponding to our point estimate.

EXAMPLE 5.29

If the null hypothesis were true, determine the chance of finding $\hat{p}$ at least as far into the tails as 0.37 under the null distribution, which is a normal distribution with mean $\mu = 0.5$ and $SE = 0.016$.

This is a normal probability problem where $x = 0.37$. First, we draw a simple graph to represent the situation, similar to what is shown in Figure 5.9. Since $\hat{p}$ is so far out in the tail, we know the tail area is going to be very small. To find it, we start by computing the Z-score using the mean of 0.5 and the standard error of 0.016:

$$Z = \frac{0.37 - 0.5}{0.016} = -8.125$$

(E)

We can use software to find the tail area: 2.2×10^{-16} (0.00000000000000022). If using the normal probability table in Appendix B.1, we'd find that $Z = -8.125$ is off the table, so we would use the smallest area listed: 0.0002.

The potential $\hat{p}$'s in the upper tail beyond 0.63, which are shown in Figure 5.10, also represent observations at least as extreme as the observed value of 0.37. To account for these values that are also more extreme under the hypothesis setup, we double the lower tail to get an estimate of the p-value: 4.4×10^{-16} (or if using the table method, 0.0004).

The p-value represents the probability of observing such an extreme sample proportion by chance, if the null hypothesis were true.

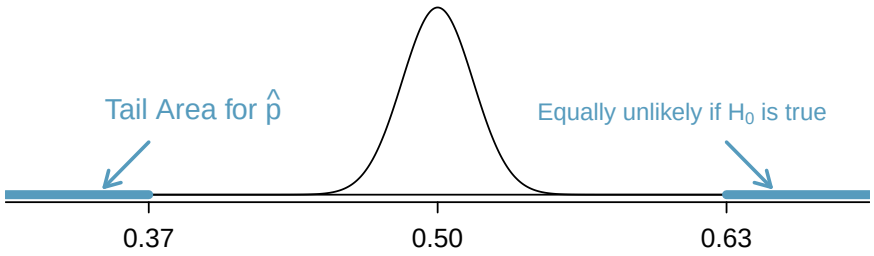


Figure 5.10: If H_0 were true, then the values above 0.63 are just as unlikely as values below 0.37.

EXAMPLE 5.30

How should we evaluate the hypotheses using the p-value of 4.4×10^{-16}? Use the standard significance level of $\alpha = 0.05$.

If the null hypothesis were true, there's only an incredibly small chance of observing such an extreme deviation of $\hat{p}$ from 0.5. This means one of the following must be true:

(E)

1. The null hypothesis is true, and we just happened to get observe something so extreme that only happens about once in every 23 quadrillion times (1 quadrillion = 1 million $\times$ 1 billion).
2. The alternative hypothesis is true, which would be consistent with observing a sample proportion far from 0.5.

The first scenario is laughably improbable, while the second scenario seems much more plausible.

Formally, when we evaluate a hypothesis test, we compare the p-value to the significance level, which in this case is $\alpha = 0.05$. Since the p-value is less than α, we reject the null hypothesis. That is, the data provide strong evidence against H_0. The data indicate the direction of the difference: a majority of Americans do not support expanding the use of coal-powered energy.

COMPARE THE P-VALUE TO α TO EVALUATE H_0

When the p-value is less than the significance level, α, reject H_0. We would report a conclusion that the data provide strong evidence supporting the alternative hypothesis.

When the p-value is greater than α, do not reject H_0, and report that we do not have sufficient evidence to reject the null hypothesis.

In either case, it is important to describe the conclusion in the context of the data.

GUIDED PRACTICE 5.31

(G) Do a majority of Americans support or oppose nuclear arms reduction? Set up hypotheses to evaluate this question.[19]

EXAMPLE 5.32

A simple random sample of 1028 US adults in March 2013 found that 56% support nuclear arms reduction. Does this provide convincing evidence that a majority of Americans supported nuclear arms reduction at the 5% significance level?

First, check conditions:

Independence. The poll was of a simple random sample of US adults, meaning the observations are independent.

Success-failure. In a one-proportion hypothesis test, this condition is checked using the null proportion, which is $p_0 = 0.5$ in this context: $np_0 = n(1 - p_0) = 1028 \times 0.5 = 514 \geq 10$.

With these conditions verified, the normal model may be applied to $\hat{p}$.

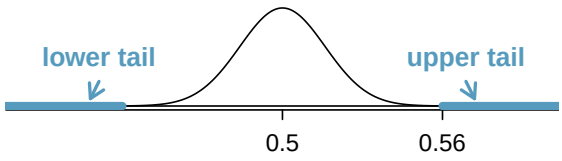
Next the standard error can be computed. The null value p_0 is used again here, because this is a hypothesis test for a single proportion.

(E)
$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1028}} = 0.0156$$

Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.56 - 0.50}{0.0156} = 3.75$$

It's generally helpful to draw null distribution and the tail areas of interest for computing the p-value:



The upper tail area is 0.0002 or less, and we double this tail area to get the p-value: 0.0004. Because the p-value is smaller than 0.05, we reject H_0. The poll provides convincing evidence that a majority of Americans supported nuclear arms reduction efforts in March 2013.

[19] We would like to understand if a majority supports or opposes, or ultimately, if there is no difference. If p is the proportion of Americans who support nuclear arms reduction, then H_0: $p = 0.50$ and H_A: $p \neq 0.50$.

HYPOTHESIS TESTING FOR A SINGLE PROPORTION

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list out hypotheses, identify the significance level, and identify $\hat{p}$ and n.

Check. Verify conditions to ensure $\hat{p}$ is nearly normal under H_0. For one-proportion hypothesis tests, use the null value to check the success-failure condition.

Calculate. If the conditions hold, compute the standard error, again using p_0, compute the Z-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

5.3.5 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 can be helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). 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.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we might choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the alternative hypothesis is actually true.

Additionally, if the cost of collecting data is small relative to the cost of a Type 2 Error, then it may also be a good strategy to collect more data. Under this strategy, the Type 2 Error can be reduced while not affecting the Type 1 Error rate. Of course, collecting extra data is often costly, so there is typically a cost-benefit analysis to be considered.

EXAMPLE 5.33

A car manufacturer is considering switching to a new, higher quality piece of equipment that constructs vehicle door hinges. They figure that they will save money in the long run if this new machine produces hinges that have flaws less than 0.2% of the time. However, if the hinges are flawed more than 0.2% of the time, they wouldn't get a good enough return-on-investment from the new piece of equipment, and they would lose money. Is there good reason to modify the significance level in such a hypothesis test?

(E)

The null hypothesis would be that the rate of flawed hinges is 0.2%, while the alternative is that it the rate is different than 0.2%. This decision is just one of many that have a marginal impact on the car and company. 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.

EXAMPLE 5.34

The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not door hinges. If the durability of these safety 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?

(E)

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

GUIDED PRACTICE 5.35

(G) A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken, so 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.[20]

WHY IS 0.05 THE DEFAULT?

The $\alpha = 0.05$ threshold is most common. But why? Maybe the standard level should be smaller, or perhaps larger. If you're a little puzzled, you're reading with an extra critical eye – good job! We've made a 5-minute task to help clarify *why 0.05*:

www.openintro.org/why05

5.3.6 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, even differences where there is no practical value. In such cases, we still say the difference is **statistically significant**, but it is not **practically significant**. For example, an online experiment might identify that placing additional ads on a movie review website statistically significantly increases viewership of a TV show by 0.001%, but this increase might not have any practical value.

One role of a data scientist in conducting a study often includes planning the size of the study. The data scientist might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain other information, such as a very rough estimate of the true proportion p, so that she could roughly estimate the standard error. From here, she can suggest a sample size that is sufficiently large that, if there is a real difference that is meaningful, we could detect it. While larger sample sizes may still be used, these calculations are especially helpful when considering costs or potential risks, such as possible health impacts to volunteers in a medical study.

5.3.7 One-sided hypothesis tests (special topic)

So far we've only considered what are called **two-sided hypothesis tests**, where we care about detecting whether p is either above or below some null value p_0. There is a second type of hypothesis

[20] 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 $\alpha = 0.01$.

test called a **one-sided hypothesis test**. For a one-sided hypothesis test, the hypotheses take one of the following forms:

1. There's only value in detecting if the population parameter is *less than* some value p_0. In this case, the alternative hypothesis is written as $p < p_0$ for some null value p_0.

2. There's only value in detecting if the population parameter is *more than* some value p_0: In this case, the alternative hypothesis is written as $p > p_0$.

While we adjust the form of the alternative hypothesis, we continue to write the null hypothesis using an equals-sign in the one-sided hypothesis test case.

In the entire hypothesis testing procedure, there is only one difference in evaluating a one-sided hypothesis test vs a two-sided hypothesis test: how to compute the p-value. In a one-sided hypothesis test, we compute the p-value as the tail area in the *direction of the alternative hypothesis only*, meaning it is represented by a single tail area. Herein lies the reason why one-sided tests are sometimes interesting: if we don't have to double the tail area to get the p-value, then the p-value is smaller and the level of evidence required to identify an interesting finding in the direction of the alternative hypothesis goes down. However, one-sided tests aren't all sunshine and rainbows: the heavy price paid is that any interesting findings in the opposite direction must be disregarded.

EXAMPLE 5.36

In Section 1.1, we encountered an example where doctors were interested in determining whether stents would help people who had a high risk of stroke. The researchers believed the stents would help. Unfortunately, the data showed the opposite: patients who received stents actually did worse. Why was using a two-sided test so important in this context?

E

Before the study, researchers had reason to believe that stents would help patients since existing research suggested stents helped in patients with heart attacks. It would surely have been tempting to use a one-sided test in this situation, and had they done this, they would have limited their ability to identify potential harm to patients.

Example 5.36 highlights that using a one-sided hypothesis creates a risk of overlooking data supporting the opposite conclusion. We could have made a similar error when reviewing the Roslings' question data this section; if we had a pre-conceived notion that college-educated folks wouldn't do worse than random guessing and so used a one-sided test, we would have missed the really interesting finding that many people have incorrect knowledge about global public health.

When might a one-sided test be appropriate to use? *Very rarely.* Should you ever find yourself considering using a one-sided test, carefully answer the following question:

> *What would I, or others, conclude if the data happens to go clearly in the opposite direction than my alternative hypothesis?*

If you or others would find any value in making a conclusion about the data that goes in the opposite direction of a one-sided test, then a two-sided hypothesis test should actually be used. These considerations can be subtle, so exercise caution. We will only apply two-sided tests in the rest of this book. [If planning to add some EOCEs that touch on this topic, then will want to tweak the text here.]

[The current plan is to cut this example. @Mine, what do you think about making this example into an odd-numbered EOCE?]

EXAMPLE 5.37

Why can't we simply run a one-sided test that goes in the direction of the data?

We've been building a careful framework that controls for the Type 1 Error, which is the significance level α in a hypothesis test. We'll use the $\alpha = 0.05$ below to keep things simple.

Imagine we could pick the one-sided test after we saw the data. What will go wrong?

E

- If $\hat{p}$ is *smaller* than the null value, then a one-sided test where $p < p_0$ would mean that any observation in the *lower* 5% tail of the null distribution would lead to us rejecting H_0.
- If $\hat{p}$ is *larger* than the null value, then a one-sided test where $p > p_0$ would mean that any observation in the *upper* 5% tail of the null distribution would lead to us rejecting H_0.

Then if H_0 were true, there's a 10% chance of being in one of the two tails, so our testing error is actually $\alpha = 0.10$, not 0.05. That is, not being careful about when to use one-sided tests effectively undermines the methods we're working so hard to develop and utilize.

Chapter 6

Inference for categorical data

In this chapter, we apply the methods and ideas from Chapter 5 in several contexts for categorical data. We'll start by revisiting what we learned for a single proportion, where the normal distribution can be used to model the uncertainty in the sample proportion. Next, we apply these same ideas to analyze the difference of two proportions using the normal model when the necessary conditions are satisfied. Later in the chapter, we apply inference techniques to contingency tables; while we will use a different distribution in this context, the core ideas of hypothesis testing remain the same.

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

6.1 Inference for a single proportion

We encountered inference methods for a single proportion in Chapter 5, exploring point estimates, confidence intervals, and hypothesis tests. In this section, we'll do a review of these topics and also explore how to perform sample size calculations for data collection purposes in the context of a single proportion.

6.1.1 Identifying when the sample proportion is nearly normal

A sample proportion $\hat{p}$ will tend to be well-modeled using a normal distribution when the observations in the sample are independent and the sample size is sufficiently large.

SAMPLING DISTRIBUTION OF $\hat{P}$

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when:

1. The sample observations are independent, e.g. are from a simple random sample.
2. We expected to see at least 10 successes and 10 failures in the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

When these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Typically we don't know the true proportion, p, so we substitute some value to check conditions and estimate the standard error. For confidence intervals, usually the sample proportion $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value – that is, the proportion claimed in the null hypothesis – is used in place of p.

6.1.2 Confidence intervals for a proportion

Often times we want to understand what a range of plausible values for the parameter, p. When this is our goal, a confidence interval is useful, which for a proportion usually takes the form of

$$\hat{p} \pm z^\star \times SE_{\hat{p}}$$

A core requirement is that $\hat{p}$ can be reasonably modeled with a normal distribution, and we check this requirement using conditions.

EXAMPLE 6.1

A simple random sample of 826 payday loan borrowers was surveyed to better understand their interests around regulation and costs. 70% of the responses supported new regulations on payday lenders. Is it reasonable to model $\hat{p} = 0.70$ using a normal distribution?

This is a random sample, so the observations are independent and representative of the population of interest.

E

We also must check the success-failure condition. In the example, we are defining a *success* as a respondent who supported new regulations and a *failure* as someone who did not. We check this condition using $\hat{p}$ in place of p when computing a confidence interval:

$$\text{Successes: } np \approx 826 * 0.70 = 578 \qquad \text{Failures: } n(1-p) \approx 826 * (1 - 0.70) = 248$$

Since both values are at least 10, we can use the normal distribution to model $\hat{p}$.

GUIDED PRACTICE 6.2

Estimate the standard error of $\hat{p} = 0.70$. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p in the formula.[1]

EXAMPLE 6.3

Construct a 95% confidence interval for p, the proportion of payday borrowers who support increased regulation for payday lenders.

Using the standard error $SE = 0.016$ from Guided Practice 6.2, the point estimate 0.70, and $z^\star = 1.96$ for a 95% confidence interval, the confidence interval is

$$\text{point estimate} \pm z^\star \times SE \quad \rightarrow \quad 0.70 \pm 1.96 \times 0.016 \quad \rightarrow \quad (0.669, 0.731)$$

We are 95% confident that the true proportion of payday borrowers who supported regulation at the time of the poll was between 0.669 and 0.731.

CONFIDENCE INTERVAL FOR A SINGLE PROPORTION

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\hat{p}$ and n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one-proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure condition.

Calculate. If the conditions hold, compute SE using $\hat{p}$, find $z^\star$, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

For additional one-proportion confidence interval examples, see Section 5.2.

6.1.3 Hypothesis testing for a proportion

One possible regulation for payday lenders is that they would be required to do a credit check and evaluate debt payments against the borrower's finances. We would like to know: would borrowers support this form of regulation?

GUIDED PRACTICE 6.4

Set up hypotheses to evaluate whether borrowers have a majority support or majority opposition for this type of regulation.[2]

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1 - p_0)$ are at least 10, where p_0 is the null value.

GUIDED PRACTICE 6.5

We consider another question asked of the payday loan borrowers: Do you support a regulation that would require lenders to pull your credit report and evaluate your debt payments? Of the 826 borrowers, 51% said they supported such a regulation. Is it reasonable to model $\hat{p} = 0.51$ for a hypothesis test here?[3]

[1] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.70(1-0.70)}{826}} = 0.016$.
[2] H_0: $p = 0.50$. H_A: $p \neq 0.50$.

EXAMPLE 6.6

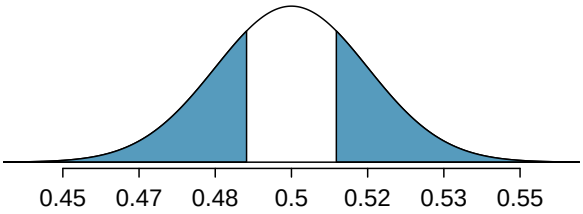
Using the hypotheses and data from Guided Practice 6.4 and 6.5, evaluate whether the poll provides convincing evidence that a majority of payday loan borrowers support a new regulation that would require lenders to pull credit reports and evaluate debt payments.

With hypotheses already set up and conditions checked, we can move onto calculations. The standard error in the context of a one-proportion hypothesis test is computed using the null value, p_0:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{826}} = 0.017$$

A picture of the normal model is shown below with the p-value represented by the shaded region.

(E)



Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.51 - 0.50}{0.017} = 0.59$$

The single tail area is 0.2776, and the p-value, represented by both tail areas together, is 0.5552. Because the p-value is larger than 0.05, we do not reject H_0. The poll does not provide convincing evidence that a majority of payday loan borrowers support or oppose regulations around credit checks and evaluation of debt payments.

HYPOTHESIS TESTING FOR A SINGLE PROPORTION

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list out hypotheses, identify the significance level, and identify $\hat{p}$ and n.

Check. Verify conditions to ensure $\hat{p}$ is nearly normal under H_0. For one-proportion hypothesis tests, use the null value to check the success-failure condition.

Calculate. If the conditions hold, compute the standard error, again using p_0, compute the Z-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

For additional one-proportion hypothesis test examples, see Section 5.3.

6.1.4 When one or more conditions aren't met

We've spent a lot of time discussing conditions for when $\hat{p}$ can be reasonably modeled by a normal distribution. What happens when the success-failure condition fails? What about when the independence condition fails? In either case, the general ideas of confidence intervals and hypothesis tests remain the same, but the strategy or technique used to generate the interval or p-value would

[3]Independence holds since the poll is based on a random sample. The success failure also holds, which we check using the null proportion, $p_0 = 0.5$ from the null hypothesis: $np \approx 826 \times 0.5 = 413$, $n(1-p) \approx 826 \times 0.5 = 413$.

change.

When the success-failure condition isn't met for a hypothesis test, we can simulate the null distribution of $\hat{p}$ using the null value, p_0. The simulation concept is similar to the ideas used in the malaria case study presented in Section 2.3, and an online section outlines this strategy:

www.openintro.org/r?go=stat_sim_prop_ht

[Port over the deleted OS3 section to be an online extra.] For a confidence interval when the success-failure condition isn't met, we can use what's called the **Clopper-Pearson interval**, where the details of this method live an internet search away, even if those details are beyond the scope of this book.

The independence condition is a more nuanced requirement. When it isn't met, it is important to understand how and why it isn't met. For example, if we took a cluster sample (see Section 1.3), suitable statistical methods are available but would be beyond the scope of even most second or third courses in statistics. On the other hand, we'd be stretched to find any method that we could confidently apply to correct the inherent biases of data from a convenience sample.

While this book is scoped to well-constrained statistical problems, do remember that this is just the first book in what is a large library of statistical methods that are suitable for a very wide range of data and contexts.

6.1.5 Choosing a sample size when estimating a proportion

When collecting data, we choose a sample size suitable for the purpose of the study. Often times this means choosing a sample size large enough that the **margin of error** – which is the part we add and subtract from the point estimate in a confidence interval – is sufficiently small that the sample is useful. More explicitly, our task is to find a sample size n so that the sample proportion is within some margin of error m of the actual proportion with a certain level of confidence.

EXAMPLE 6.7

A university newspaper is conducting a survey to determine what fraction of students support a $200 per year increase in fees to pay for a new football stadium. How big of a sample is required to ensure the margin of error is smaller than 0.04 using a 95% confidence level?

The margin of error for a sample proportion is

$$z^\star \sqrt{\frac{p(1-p)}{n}}$$

Our goal is to find the smallest sample size n so that this margin of error is smaller than $m = 0.04$. For a 95% confidence level, the value $z^\star$ corresponds to 1.96:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \ < \ 0.04$$

(E)

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could enter in that value and solve for n. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this *worst case value* if no estimate of the proportion is available:

$$\begin{aligned} 1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} \ &< \ 0.04 \\ 1.96^2 \times \frac{0.5(1-0.5)}{n} \ &< \ 0.04^2 \\ 1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} \ &< \ n \\ 600.25 \ &< \ n \end{aligned}$$

We would need over 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

When an estimate of the proportion is available, we use it in place of the worst case proportion value, 0.5.

GUIDED PRACTICE 6.8

(G)

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 1% with a 90% confidence level. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.[4]

---

[4] For a 90% confidence interval, $z^\star = 1.65$, and since an estimate of the proportion 0.017 is available, we'll use it in the margin of error formula:

$$1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \ < \ 0.01 \quad \rightarrow \quad \frac{0.017(1-0.017)}{n} \ < \ \left(\frac{0.01}{1.65}\right)^2 \quad \rightarrow \quad 454.96 \ < \ n$$

For sample size calculations, we always round up, so the first tire model suggests 455 tires would be sufficient.

A similar computation can be accomplished using 0.062 and 0.013 for p, and you should verify that using these proportions results in minimum sample sizes of 1584 and 350 tires, respectively.

EXAMPLE 6.9

The sample sizes vary widely in Guided Practice 6.8. Which of the three would you suggest using? What would influence your choice?

E

We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. There are also other reasonable approaches.

Also observe that the success-failure condition would need to be checked in the final sample. For instance, if we sampled $n = 1584$ tires and found a failure rate of 0.5%, the normal approximation would not be reasonable, and we would require more advanced statistical methods for creating the confidence interval.

GUIDED PRACTICE 6.10

G

Suppose we want to continually track the support of payday borrowers for regulation on lenders, where we would conduct a new poll every month. Running such frequent polls is expensive, so we decide a wider margin of error of 5% for each individual survey would be acceptable. Based on the original sample of borrowers where 70% supported some form of regulation, how big should our monthly sample be for a margin of error of 0.04 with 95% confidence?[5]

[5] We complete the same computations as before, except now we use 0.70 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.70(1-0.70)}{n}} \leq 0.05 \qquad \rightarrow \qquad n \geq 322.7$$

A sample size of 323 or more would be reasonable. (Reminder: always round up for sample size calculations!) Given that we plan to track this poll over time, we also may want to periodically repeat these calculations to ensure that we're being thoughtful in our sample size recommendations.

6.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the utility of a blood thinner for heart attack patients. In the second application, we examine the efficacy of mammograms in reducing deaths from breast cancer. In the last example, a quadcopter company weighs whether to switch to a higher quality manufacturer of rotor blades.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

6.2.1 Sample distribution of the difference of two proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

CONDITIONS FOR THE SAMPLING DISTRIBUTION OF $\hat{P}_1 - \hat{P}_2$ TO BE NORMAL

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when

- each proportion separately follows a normal model, and
- the two samples are independent of each other.

The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

6.2.2 Confidence intervals for $p_1 - p_2$

In the setting of confidence intervals for a difference of two proportions, the two sample proportions are used to verify the success-failure condition and also compute the standard error, just as was the case with a single proportion. The generic formula of a confidence interval will apply in this case, just as it did in the one-proportion case:

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

As with a single proportion context, we can follow the same Prepare, Check, Calculate, Conclude steps for computing a confidence interval or completing a hypothesis test for a difference of two proportions. The details change a little, but the general approach remain the same. Think about these steps when you apply statistical methods.

EXAMPLE 6.11

We consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital. These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours. The results are shown in Figure 6.1. Create and interpret a 90% confidence interval of the difference in the survival rates.

First the conditions must be verified. Because the patients were randomized to their groups and one heart attack patient is unlikely to influence the next that was in the study, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each group, where we have at least 10 successes and 10 failures in each experiment arm. With these conditions met, the normal model can be used for the point estimate of the difference in survival rates, where we'll use p_t corresponds to the survival rate in the treatment group and p_c for the control group:

(E)

$$\hat{p}_t - \hat{p}_c = \frac{14}{40} - \frac{11}{50} = 0.35 - 0.22 = 0.13$$

The standard error may be computed from We use the standard error formula provided on page 163. As with the one-sample proportion case, we use our best estimates of each proportion in the formula:

$$SE \approx \sqrt{\frac{0.35(1-0.35)}{40} + \frac{0.22(1-0.22)}{50}} = 0.095$$

For a 90% confidence interval, we use $z^\star = 1.65$:

$$\text{point estimate} \ \pm \ z^\star \times SE \quad \rightarrow \quad 0.13 \ \pm \ 1.65 \times 0.095 \quad \rightarrow \quad (-0.027, 0.287)$$

We are 90% confident that that blood thinners have a difference of -2.7% to +28.7% percentage point impact on survival rate for patients like those in the study. That is, we do not have enough information to say with confidence whether blood thinners help or harm heart attack patients who have been admitted after they have undergone CPR.

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

Figure 6.1: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

GUIDED PRACTICE 6.12

A 5-year experiment was conducted to evaluate the effectiveness of fish oils on reducing cardiovascular events, where each subject was randomized into one of two treatment groups. We'll consider heart attack outcomes in these patients:

(G)

	heart_attack	no_event	Total
fish_oil	145	12788	12933
placebo	200	12738	12938

Create a 95% confidence interval for the effect of fish oils on heart attacks for patients who are well-represented by those in the study. Also interpret the interval in the context of the study.[6]

[6] The experiment randomized, subjects are unlikely to interact with each other, the observations are independent. The success-failure condition is also met for both groups as all counts are at least 10. Compute the sample proportions ($\hat{p}_{\text{fish oil}} = 0.0112$, $\hat{p}_{\text{placebo}} = 0.0155$), point estimate of the difference ($0.0112 - 0.0155 = -0.0043$), and standard error

6.2.3 Hypothesis tests for $p_1 - p_2$

A mammogram is an X-ray procedure used to check for breast cancer. Whether mammograms should be used is part of a controversial discussion, and it's the topic of our next example where we examine 2-proportion hypothesis test when H_0 is $p_1 - p_2 = 0$ (or equivalently, $p_1 = p_2$).

A 30-year study was conducted with nearly 90,000 female participants. During a 5-year screening period, each woman was randomized to one of two groups: in the first group, women received regular mammograms to screen for breast cancer, and in the second group, women received regular non-mammogram breast cancer exams. No intervention was made during the following 25 years of the study, and we'll consider death resulting from breast cancer over the full 30-year period. Results from the study are summarized in Figure 6.2.

If mammograms are much more effective than non-mammogram breast cancer exams, then we would expect to see additional deaths from breast cancer in the control group. On the other hand, if mammograms are not as effective as regular breast cancer exams, we would expect to see an increase in breast cancer deaths in the mammogram group.

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

Figure 6.2: Summary results for breast cancer study.

GUIDED PRACTICE 6.13

Is this study an experiment or an observational study?[7]

GUIDED PRACTICE 6.14

Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups.[8]

In Example 6.15, we will check the conditions for using the normal model to analyze the results of the study. The details are very similar to that of confidence intervals. However, this time we use a special proportion called the **pooled proportion** to check the success-failure condition:

$$\begin{aligned}\hat{p}_{pooled} &= \frac{\text{\# of patients who died from breast cancer in the entire study}}{\text{\# of patients in the entire study}} \\ &= \frac{500 + 505}{500 + 44{,}425 + 505 + 44{,}405} \\ &= 0.0112\end{aligned}$$

This proportion is an estimate of the breast cancer death rate across the entire study, and it's our best estimate of the proportions p_{mgm} and p_{ctrl} *if the null hypothesis is true that* $p_{mgm} = p_{ctrl}$. We will also use this pooled proportion when computing the standard error.

($SE = \sqrt{\frac{0.0112 \times 0.9888}{12933} + \frac{0.0155 \times 0.9845}{12938}} = 0.00145$). Next, plug the values into the general formula for a confidence interval, where we'll use a 95% confidence level with $z^\star = 1.96$:

$$-0.0043 \pm 1.96 \times 0.00145 \quad \rightarrow \quad (-0.0071, -0.0015)$$

We are 95% confident that fish oils decreases heart attacks by 0.15 to 0.71 percentage points (off of a baseline of about 1.55%) over a 5-year period for subjects who are similar to those in the study. Because the interval is entirely below 0, the data provide strong evidence that fish oil supplements reduce heart attacks in patients like those in the study.

[7] This is an experiment. Patients were randomized to receive mammograms or a standard breast cancer exam. We will be able to make causal conclusions based on this study.

[8] H_0: the breast cancer death rate for patients screened using mammograms is the same as the breast cancer death rate for patients in the control, $p_{mgm} - p_{ctrl} = 0$.
H_A: the breast cancer death rate for patients screened using mammograms is different than the breast cancer death rate for patients in the control, $p_{mgm} - p_{ctrl} \neq 0$.

EXAMPLE 6.15

Can we use the normal model to analyze this study?

Because the patients are randomized, they can be treated as independent.

We also must check the success-failure condition for each group. Under the null hypothesis, the proportions p_{mgm} and p_{ctrl} are equal, so we check the success-failure condition with our best estimate of these values under H_0, the pooled proportion from the two samples, $\hat{p}_{pooled} = 0.0112$:

$$\hat{p}_{pooled} \times n_{mgm} = 0.0112 \times 44{,}925 = 503 \qquad (1 - \hat{p}_{pooled}) \times n_{mgm} = 0.9888 \times 44{,}925 = 44{,}422$$
$$\hat{p}_{pooled} \times n_{ctrl} = 0.0112 \times 44{,}910 = 503 \qquad (1 - \hat{p}_{pooled}) \times n_{ctrl} = 0.9888 \times 44{,}910 = 44{,}407$$

The success-failure condition is satisfied since all values are at least 10, and we can safely apply the normal model.

USE THE POOLED PROPORTION ESTIMATE WHEN H_0 IS $P_1 - P_2 = 0$

When the null hypothesis is that the proportions are equal, use the pooled proportion ($\hat{p}_{pooled}$) to verify the success-failure condition and estimate the standard error:

$$\hat{p}_{pooled} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

In Example 6.15, the pooled proportion was used to check the success-failure condition.[9] In the next example, we see the second place where the pooled proportion comes into play: the standard error calculation.

EXAMPLE 6.16

Compute the point estimate of the difference in breast cancer death rates in the two groups, and use the pooled proportion $\hat{p}_{pooled} = 0.0112$ to calculate the standard error.

The point estimate of the difference in breast cancer death rates is

$$\begin{aligned} \hat{p}_{mgm} - \hat{p}_{ctrl} &= \frac{500}{500 + 44{,}425} - \frac{505}{505 + 44{,}405} \\ &= 0.01113 - 0.01125 \\ &= -0.00012 \end{aligned}$$

The breast cancer death rate in the mammogram group was 0.012% less than in the control group. Next, the standard error is calculated *using the pooled proportion*, $\hat{p}_{pooled}$:

$$SE = \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{mgm}} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{ctrl}}} = 0.00070$$

[9] For an example of a two-proportion hypothesis test that does not require the the success-failure condition to be met, see Section 2.3.

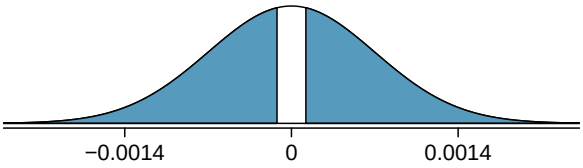
EXAMPLE 6.17

Using the point estimate $\hat{p}_{mgm} - \hat{p}_{ctrl} = -0.00012$ and standard error $SE = 0.00070$, calculate a p-value for the hypothesis test and write a conclusion.

Just like in past tests, we first compute a test statistic and draw a picture:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{-0.00012 - 0}{0.00070} = -0.17$$

(E)



The lower tail area is 0.4325, which we double to get the p-value: 0.8650. Because this p-value is larger than 0.05, we do not reject the null hypothesis. That is, the difference in breast cancer death rates is reasonably explained by chance, and we do not observe benefits or harm from mammograms relative to a regular breast exam.

Can we conclude that mammograms have no benefits or harm? Here are a few important considerations to keep in mind when reviewing the mammogram study as well as any other medical study:

- If mammograms are helpful or harmful, the data suggest the effect isn't very large. So while we do not accept the null hypothesis, we also don't have sufficient evidence to conclude that mammograms reduce or increase breast cancer deaths.
- Are mammograms more or less expensive than a non-mammogram breast exam? If one option is much more expensive than the other and doesn't offer clear benefits, then we should lean towards the less expensive option.
- The study's authors also found that mammograms led to overdiagnosis of breast cancer, which means some breast cancers were found (or thought to be found) but that these cancers would not cause symptoms during patients' lifetimes. That is, something else would kill the patient before breast cancer symptoms appeared. This means some patients may have been treated for breast cancer unnecessarily, and this treatment is another cost to consider. It is also important to recognize that overdiagnosis can cause unnecessary physical or emotional harm to patients.

These considerations highlight the complexity around medical care and treatment recommendations. Experts and medical boards who study medical treatments use considerations like those above to provide their best recommendation based on the current evidence.

6.2.4 More on 2-proportion hypothesis tests (special topic)

When we conduct a 2-proportion hypothesis test, usually H_0 is $p_1 - p_2 = 0$. However, there are rare situations where we want to check for some difference in p_1 and p_2 that is some value other than 0. For example, maybe we care about checking a null hypothesis where $p_1 - p_2 = 0.1$.[10] In contexts like these, we generally use $\hat{p}_1$ and $\hat{p}_2$ to check the success-failure condition and construct the standard error.

[10] We can also encounter a similar situation with a difference of two means, though no such example is given in Chapter 7 since the methods remain exactly the same in the context of sample means. On the other hand, the success-failure condition and the calculation of the standard error vary slightly in different proportion contexts.

Figure 6.3: A Phantom quadcopter.

Photo by David J (http://flic.kr/p/oiWLNu). CC-BY 2.0 license.
This photo has been cropped and a border has been added.

GUIDED PRACTICE 6.18

A quadcopter company is considering a new manufacturer for rotor blades. The new manufacturer would be more expensive but their higher-quality blades are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive blades are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of blades that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[11]

[11] H_0: The higher-quality blades will pass inspection just 3% more frequently than the standard-quality blades. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher-quality blades will pass inspection >3% more often than the standard-quality blades. $p_{highQ} - p_{standard} > 0.03$.

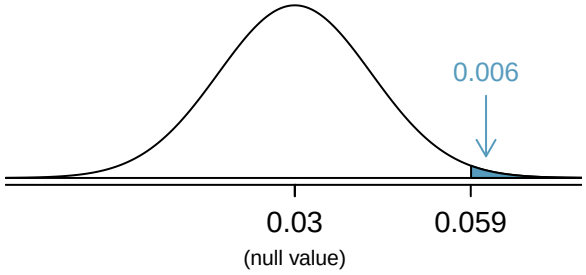


Figure 6.4: Distribution of the test statistic if the null hypothesis was true. The p-value is represented by the shaded area.

EXAMPLE 6.19

The quality control engineer from Guided Practice 6.18 collects a sample of blades, examining 1000 blades from each company and finds that 899 blades pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Guided Practice 6.18 with a significance level of 5%.

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the blades are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, $0.958 - 0.899 = 0.059$, can be said to come from a nearly normal distribution.

The standard error is computed using the two sample proportions since we do not use a pooled proportion for this context:

(E)

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114$$

In this hypothesis test, because the null is that $p_1 - p_2 = 0.03$, the sample proportions were used for the standard error calculation rather than a pooled proportion.

Next, we compute the test statistic and use it to find the p-value, which is depicted in Figure 6.4.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher-quality blades actually do pass inspection more than 3% as often as the currently used blades. Based on these results, management will approve the switch to the new supplier.

6.2.5 Examining the standard error formula (special topic)

The formula for the standard error of the difference in two proportions is similar to the formula for other standard errors. Recall that the standard error of a single proportion, $\hat{p}_1$, is

$$SE_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}}$$

The standard error of the difference of two sample proportions can be constructed from the standard errors of the separate sample proportions:

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

This special relationship follows from probability theory.

GUIDED PRACTICE 6.20

Prerequisite: Section 3.4. We can rewrite the equation above in a different way:

$$SE_{\hat{p}_1-\hat{p}_2}^2 = SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$$

Explain where this formula comes from using the ideas of probability theory.[12]

[12] The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances σ_x^2 and σ_y^2, then the variance of $X - Y$ is $\sigma_x^2 + \sigma_y^2$. Likewise, the variance corresponding to $\hat{p}_1 - \hat{p}_2$ is $\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2$. Because $\sigma_{\hat{p}_1}^2$ and $\sigma_{\hat{p}_2}^2$ are just another way of writing $SE_{\hat{p}_1}^2$ and $SE_{\hat{p}_2}^2$, the variance associated with $\hat{p}_1 - \hat{p}_2$ may be written as $SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$.

6.3 Testing for goodness of fit using chi-square

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test.

In the first case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Figure 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reflect the population of eligible jurors, i.e. registered voters.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Figure 6.5: Representation by race in a city's juries and population.

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

A second application, assessing the fit of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for 25 years are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

6.3.1 Creating a test statistic for one-way tables

EXAMPLE 6.21

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

(E)

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

GUIDED PRACTICE 6.22

(G) Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race? Answers can be found in Figure 6.6.

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Figure 6.6: Actual and expected make-up of the jurors.

We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

6.3.2 The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

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

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to first compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

Black

$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

Hispanic

$$Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39$$

Other

$$Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they – as a group – tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58$$

[13] Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown – though not here – that the square root of the count is a better way to standardize the count differences.

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual – e.g. a standardized difference of 2.5 – will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \cdots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4}$$

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a *chi-square distribution*. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

6.3.3 The chi-square distribution and finding areas

The **chi-square distribution** is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

GUIDED PRACTICE 6.23

(G) Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

Figure 6.7 and Guided Practice 6.23 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inflates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. The most common ways to do this are using computer software, using a graphing calculator, or using a table. For folks wanting to use the table option, we provide an outline of how to read the chi-square table in Appendix B.3, which is also where you may find the table. [If giving some **R** in the text, then put **R** code in the examples / exercises below.] For the examples below, use your preferred approach to confirm you get the same answers.

EXAMPLE 6.24

(E) Figure 6.8(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Find the shaded area.

Using statistical software or a graphing calculator, we can find that the upper tail area for a chi-square distribution with 3 degrees of freedom (df) and a cutoff of 6.25 is 0.1001. That is, the shaded upper tail of Figure 6.8(a) has area 0.1.

[14] (a) The center becomes larger. If took a careful look, we could see that the mean of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for $df = 2$, and then the distributions become more symmetric for the larger degrees of freedom $df = 4$ and $df = 9$. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

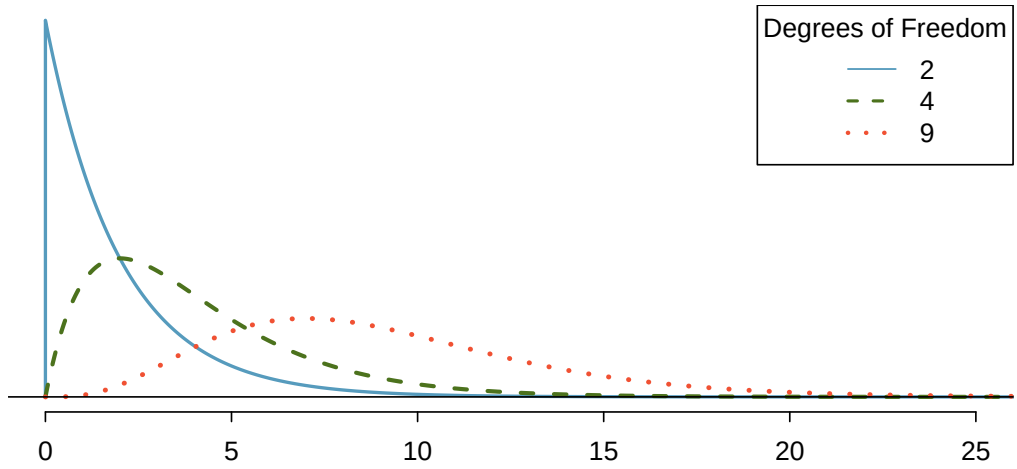


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

EXAMPLE 6.25

Figure 6.8(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3. Find the tail area.

Using software, we can find that the tail area shaded in Figure 6.8(b) to be 0.1165. If using a table, we would only be able to find a range of values for the tail area: between 0.1 and 0.2.

EXAMPLE 6.26

Figure 6.8(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Using software, we would obtain a tail area of 0.4038. If using the table in Appendix B.3, we would have identified that the tail area is larger than 0.3 but not be able to give the precise value.

GUIDED PRACTICE 6.27

Figure 6.8(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

GUIDED PRACTICE 6.28

Figure 6.8(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

GUIDED PRACTICE 6.29

Figure 6.8(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

[15] The area is 0.1109. If using a table, we would identify that it falls between 0.1 and 0.2.
[16] Precise value: 0.0404. If using the table: between 0.02 and 0.05.
[17] Precise value: 0.0266. If using the table: between 0.02 and 0.05.

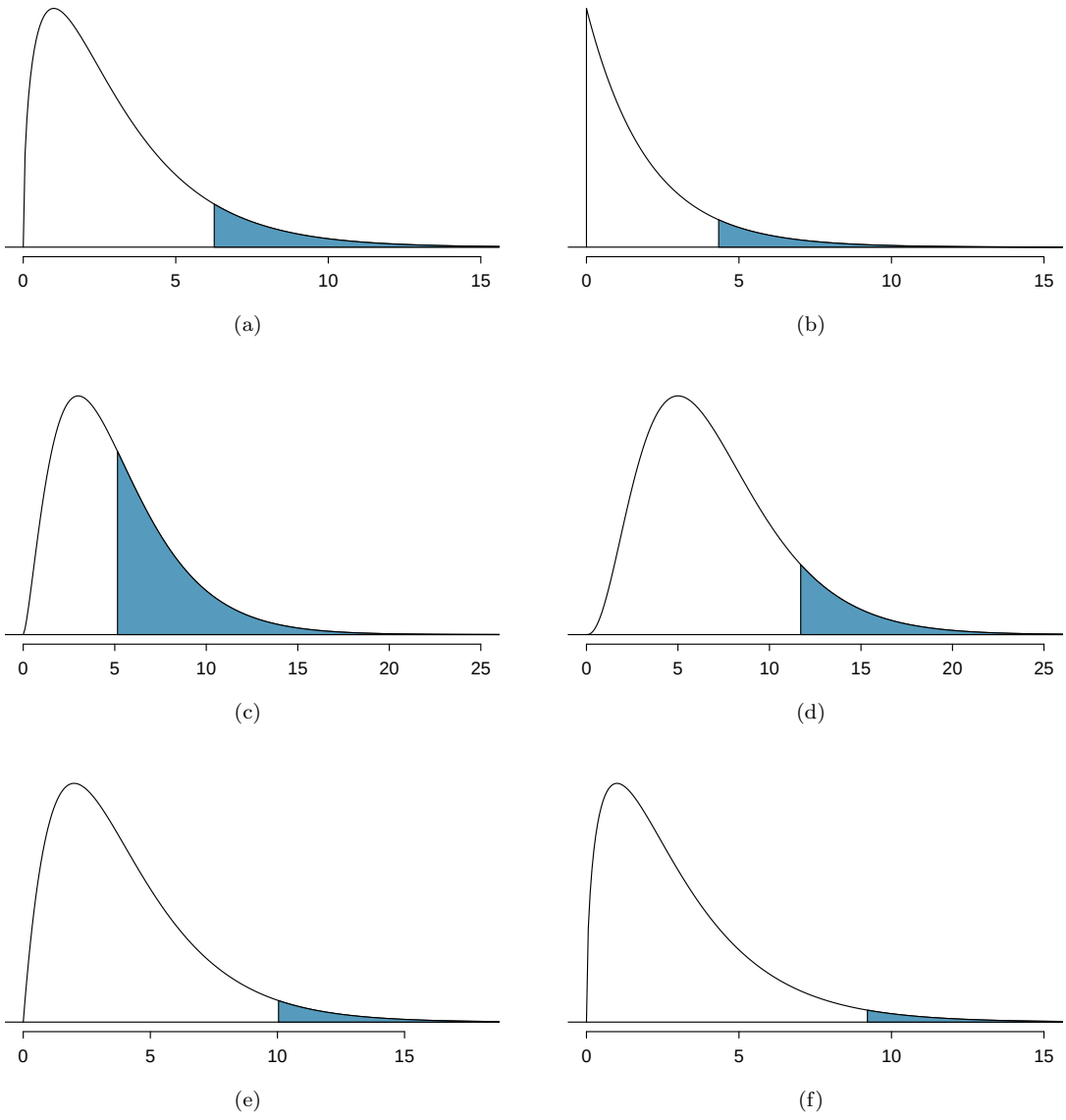


Figure 6.8: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

6.3.4 Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of bins.

EXAMPLE 6.30

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

(E)

In the jurors example, there were $k = 4$ categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with $k - 1 = 3$ degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

EXAMPLE 6.31

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

(E)

The chi-square distribution and p-value are shown in Figure 6.9. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using statistical software (or the table in Appendix B.3), we can determine that the area is 0.1171. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

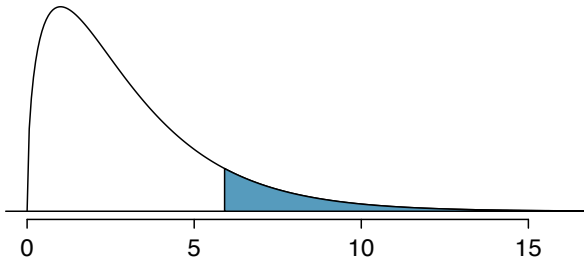


Figure 6.9: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

CHI-SQUARE TEST FOR ONE-WAY TABLE

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts O_1, O_2, ..., O_k in k categories are unusually different from what might be expected under a null hypothesis. Call the *expected counts* that are based on the null hypothesis E_1, E_2, ..., E_k. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with $k - 1$ degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

CONDITIONS FOR THE CHI-SQUARE TEST

There are two conditions that must be checked before performing a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one-proportion methods introduced in Section 6.1.

6.3.5 Evaluating goodness of fit for a distribution

Section 4.2 would be useful background reading for this example, but it is not a prerequisite.

We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model fits a data set. Daily stock returns from the S&P500 for 10 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as `Up` or `Down` (`D`) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each `Up` day:

Change in price	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
Outcome	Up	D	Up	D	Up	Up	D	D	D	Up
Days to Up	1	-	2	-	2	1	-	-	-	4

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the k^{th} trial to observe the first success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next `Up` trading day, and two more for the third `Up` day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Figure 6.10 shows the number of waiting days for a positive trading day during 10 years for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362

Figure 6.10: Observed distribution of the waiting time until a positive trading day for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362
Geometric Model	743	338	154	70	32	14	12	1362

Figure 6.11: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model ($P(D) = (1 - 0.545)^{D-1}(0.545)$) and multiply by the total number of streaks, 1362. For example, waiting for three days occurs under the geometric model about $0.455^2 \times 0.545 = 11.28\%$ of the time, which corresponds to $0.1128 \times 1362 = 154$ streaks.

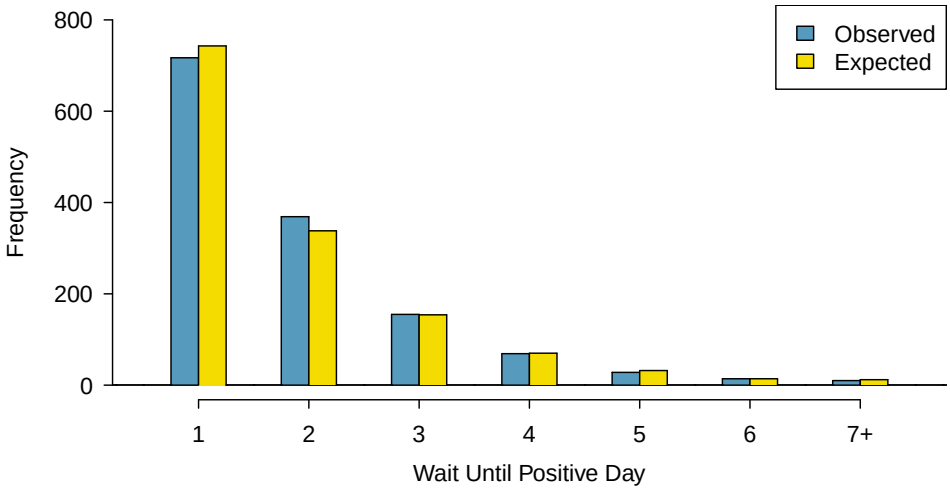


Figure 6.12: Side-by-side bar plot of the observed and expected counts for each waiting time.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock index. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a *geometric distribution*. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an `Up` day is observed. Under this hypothesis, the number of days until an `Up` day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an `Up` day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 and summarize the waiting times in Figure 6.11 and Figure 6.12. The S&P500 was positive on 54.5% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have *binned* together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the *Observed* row in Figure 6.11, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Figure 6.11. In general, the expected counts are determined by (1) identifying the null proportion associated with each bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

EXAMPLE 6.32

Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

(E)

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Figure 6.11.

GUIDED PRACTICE 6.33

(G) Figure 6.11 provides a set of count data for waiting times ($O_1 = 717$, $O_2 = 369$, ...) and expected counts under the geometric distribution ($E_1 = 743$, $E_2 = 338$, ...). Compute the chi-square test statistic, X^2.[18]

GUIDED PRACTICE 6.34

(G) Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

EXAMPLE 6.35

If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 4.61$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

(E)

Figure 6.13 shows the chi-square distribution, cutoff, and the shaded p-value. Using software, we can find the p-value: 0.5951. Ultimately, we do not have sufficient evidence to reject the notion that the wait times follow a geometric distribution for the last 10 years of data for the S&P500, i.e. we cannot reject the notion that trading days are independent.

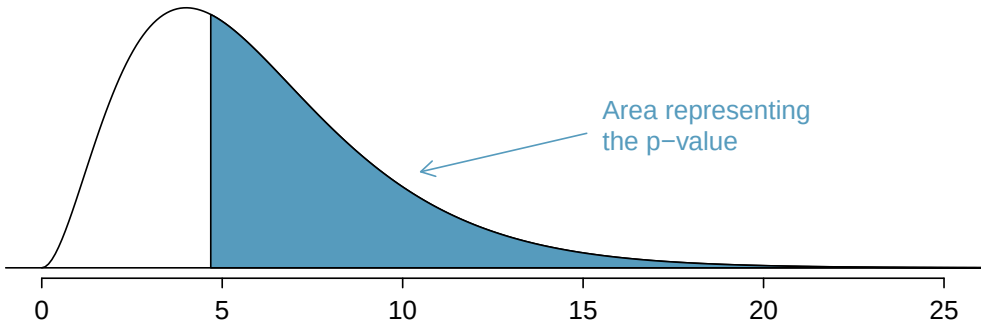


Figure 6.13: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

[18] $X^2 = \frac{(717-743)^2}{743} + \frac{(369-338)^2}{338} + \cdots + \frac{(10-12)^2}{12} = 4.61$

[19] There are $k = 7$ groups, so we use $df = k - 1 = 6$.

EXAMPLE 6.36

In Example 6.35, we did not reject the null hypothesis that the trading days are independent during the last 10 of data. Why is this so important?

(E)

It may be tempting to think the market is "due" for an `Up` day if there have been several consecutive days where it has been down. However, we haven't found strong evidence that there's any such property where the market is "due" for a correction. At the very least, the analysis suggests any dependence between days is very weak.

6.4 Testing for independence in two-way tables

We all buy used products – cars, computers, textbooks, and so on – and we sometimes assume the sellers of those products will be forthright about any underlying problems with what they're selling. This is not something we should take for granted. Researchers recruited 219 participants in a study where they would sell a used iPod[20] that was known to have frozen twice in the past. The participants were incentivized to get as much money as they could for the iPod since they would receive a 5% cut of the sale on top of $10 for participating. The researchers wanted to understand what types of questions would elicit the seller to disclose the freezing issue.

Unbeknownst to the participants who were the sellers in the study, the buyers were collaborating with the researchers to evaluate the influence of different questions on the likelihood of getting the sellers to disclose the past issues with the iPod. The scripted buyers started with "Okay, I guess I'm supposed to go first. So you've had the iPod for 2 years ..." and ended with one of three questions:

- General: What can you tell me about it?
- Positive Assumption: It doesn't have any problems, does it?
- Negative Assumption: What problems does it have?

The question is the treatment given to the sellers, and the response is whether the question prompted them to disclose the freezing issue with the iPod. The results are shown in Figure 6.14, and the data suggest that asking the, *What problems does it have?*, was the most effective at getting the seller to disclose the past freezing issues. However, you should also be asking yourself: could we see these results due to chance alone, or is this in fact evidence that some questions are more effective for getting at the truth?

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

Figure 6.14: Summary of the iPod study, where a question was posed to the study participant who acted

DIFFERENCES OF ONE-WAY TABLES VS TWO-WAY TABLES

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for *combinations* of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for the iPod experiment is really about assessing whether there is statistically significant evidence that the success each question had on getting the participant to disclose the problem with the iPod. In other words, the goal is to check whether the buyer's question was independent of whether the seller disclosed a problem.

[20] For readers not as old as the authors, an iPod is basically an iPhone without any cellular service, assuming it was one of the later generations. Earlier generations were more basic.

6.4.1 Expected counts in two-way tables

EXAMPLE 6.37

From the experiment, we can compute the proportion of all sellers who disclosed the freezing problem as $61/219 = 0.2785$. If there really is no difference among the questions and 27.85% of sellers were going to disclose the freezing problem no matter the question that was put to them, how many of the 73 people in the `General` group would we have expected to disclose the freezing problem?

E

We would predict that $0.2785 \times 73 = 20.33$ sellers would disclose the problem. Obviously we observed fewer than this, though it is not yet clear if that is due to chance variation or whether that is because the questions vary in how effective they are at getting to the truth.

GUIDED PRACTICE 6.38

G If the questions were actually equally effective, meaning about 27.85% of respondents would disclose the freezing issue regardless of what question they were asked, about how many sellers would we expect to *hide* the freezing problem from the Positive Assumption group?[21]

We can compute the expected number of sellers who we would expect to disclose or hide the freezing issue for all groups, if the questions had no impact on what they disclosed, using the same strategy employed in Example 6.37 and Guided Practice 6.38. These expected counts were used to construct Figure 6.15, which is the same as Figure 6.14, except now the expected counts have been added in parentheses.

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2 **(20.33)**	23 **(20.33)**	36 **(20.33)**	61
Hide Problem	71 **(52.67)**	50 **(52.67)**	37 **(52.67)**	158
Total	73	73	73	219

Figure 6.15: The observed counts and the **(expected counts)**.

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 27.85% of each column should be in the first row:

$$0.2785 \times (\text{column 1 total}) = 20.33$$
$$0.2785 \times (\text{column 2 total}) = 20.33$$
$$0.2785 \times (\text{column 3 total}) = 20.33$$

Looking back to how 0.2785 was computed – as the fraction of sellers who disclosed the freezing issue (158/219) – these three expected counts could have been computed as

$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) = 20.33$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

[21] We would expect $(1 - 0.2785) \times 73 = 52.67$. It is okay that this result, like the result from Example 6.37, is a fraction.

COMPUTING EXPECTED COUNTS IN A TWO-WAY TABLE

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i,\text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

6.4.2 The chi-square test for two-way tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

General formula	$\dfrac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\dfrac{(2-20.33)^2}{20.33} = 16.53$
Row 1, Col 2	$\dfrac{(23-20.33)^2}{20.33} = 0.35$
$\vdots$	$\vdots$
Row 2, Col 3	$\dfrac{(37-52.67)^2}{52.67} = 4.66$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 16.53 + 0.35 + \cdots + 4.66 = 40.13$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1})$$

In our example, the degrees of freedom parameter is

$$df = (2-1) \times (3-1) = 2$$

If the null hypothesis is true (i.e. the questions had no impact on the sellers in the experiment), then the test statistic $X^2 = 40.13$ closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.16.

COMPUTING DEGREES OF FREEDOM FOR A TWO-WAY TABLE

When applying the chi-square test to a two-way table, we use

$$df = (R-1) \times (C-1)$$

where R is the number of rows in the table and C is the number of columns.

When analyzing 2-by-2 contingency tables, one guideline is to use the two-proportion methods introduced in Section 6.2.

[22] Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.

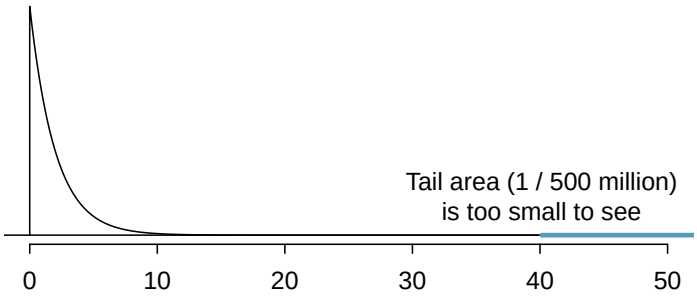


Figure 6.16: Visualization of the p-value for $X^2 = 40.13$ when $df = 2$.

EXAMPLE 6.39

Find the p-value and draw a conclusion about whether the question affects the sellers likelihood of reporting the freezing problem.

(E)

Using a computer, we can compute a very precise value for the tail area above $X^2 = 40.13$ for a chi-square distribution with 2 degrees of freedom: 0.000000002. (If using the table in Appendix B.3, we would identify the p-value is smaller than 0.001.) Using a significance level of $\alpha = 0.05$, the null hypothesis is rejected since the p-value is smaller. That is, the data provide convincing evidence that the question asked did affect a seller's likelihood to tell the truth about problems with the iPod.

EXAMPLE 6.40

Figure 6.17 summarizes the results of an experiment evaluating three treatments for Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The three treatments considered were continued treatment with metformin (`met`), treatment with metformin combined with rosiglitazone (`rosi`), or a lifestyle intervention program. Each patient had a primary outcome, which was either lacked glycemic control (failure) or did not lack that control (success). What are appropriate hypotheses for this test?

(E)

H_0: There is no difference in the effectiveness of the three treatments.

H_A: There is some difference in effectiveness between the three treatments, e.g. perhaps the `rosi` treatment performed better than `lifestyle`.

	Failure	Success	Total
`lifestyle`	109	125	234
`met`	120	112	232
`rosi`	90	143	233
Total	319	380	699

Figure 6.17: Results for the Type 2 Diabetes study.

GUIDED PRACTICE 6.41

(G)

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.40. As a first step, compute the expected values for each of the six table cells.[23]

[23] The expected count for row one / column one is found by multiplying the row one total (234) and column one total (319), then dividing by the table total (699): $\frac{234 \times 319}{699} = 106.8$. Similarly for the second column and the first row: $\frac{234 \times 380}{699} = 127.2$. Row 2: 105.9 and 126.1. Row 3: 106.3 and 126.7.

GUIDED PRACTICE 6.42

Compute the chi-square test statistic for the data in Figure 6.17.[24]

GUIDED PRACTICE 6.43

Because there are 3 rows and 2 columns, the degrees of freedom for the test is $df = (3-1)\times(2-1) = 2$. Use $X^2 = 8.16$, $df = 2$, evaluate whether to reject the null hypothesis using a significance level of 0.05.[25]

[24] For each cell, compute $\frac{(\text{obs}-\text{exp})^2}{exp}$. For instance, the first row and first column: $\frac{(109-106.8)^2}{106.8} = 0.05$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 0.05 + \cdots + 2.11 = 8.16$.

[25] If using a computer, we can identify the p-value as 0.017. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that at least one of the treatments is more or less effective than the others at treating Type 2 Diabetes for glycemic control.

Chapter 7

Inference for numerical data

Chapters 5 introduced a framework for statistical inference based on confidence intervals and hypotheses using the normal distribution. In this chapter, we encounter several new point estimates and a couple new distributions. In each case, the inference ideas remain the same: determine which point estimate or test statistic is useful, identify an appropriate distribution for the point estimate or test statistic, and apply the ideas from Chapter 5.

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

7.1 One-sample means with the t-distribution

We required a large sample in Chapter 5 for two reasons:

1. The sampling distribution of $\bar{x}$ tends to be more normal when the sample is large.
2. The calculated standard error is typically very accurate when using a large sample.

So what should we do when the sample size is small? As we'll discuss in Section 7.1.1, if the population data are nearly normal, then $\bar{x}$ will also follow a normal distribution, which addresses the first problem. The accuracy of the standard error is trickier, and for this challenge we'll introduce a new distribution called the t-distribution.

While we emphasize the use of the t-distribution for small samples, this distribution is also generally used for large samples, where it produces similar results to those from the normal distribution.

7.1.1 The normality condition

A special case of the Central Limit Theorem ensures the sample mean will be nearly normal, regardless of sample size, when the data come from a nearly normal distribution.

CENTRAL LIMIT THEOREM FOR NORMAL DATA

The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

7.1.2 Introducing the t-distribution

In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t-distribution. A t-distribution, shown as a solid line in Figure 7.1, has a bell shape. However, its tails are thicker than the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[1] While our estimate of the standard error will be a little less accurate when we are analyzing a small data set, these extra thick tails of the t-distribution are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t-distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell-shaped t-distribution. Several t-distributions are shown in Figure 7.2. When there are more degrees of freedom, the t-distribution looks very much like the standard normal distribution.

DEGREES OF FREEDOM (DF)

The degrees of freedom describe the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

[1] The standard deviation of the t-distribution is actually a little more than 1. However, it is useful to always think of the t-distribution as having a standard deviation of 1 in all of our applications.

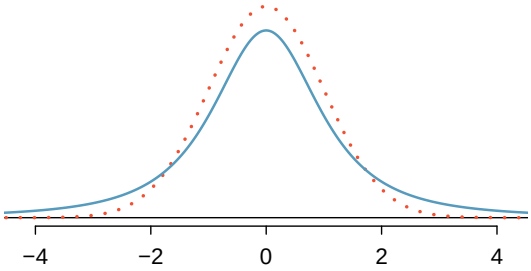


Figure 7.1: Comparison of a t-distribution (solid line) and a normal distribution (dotted line).

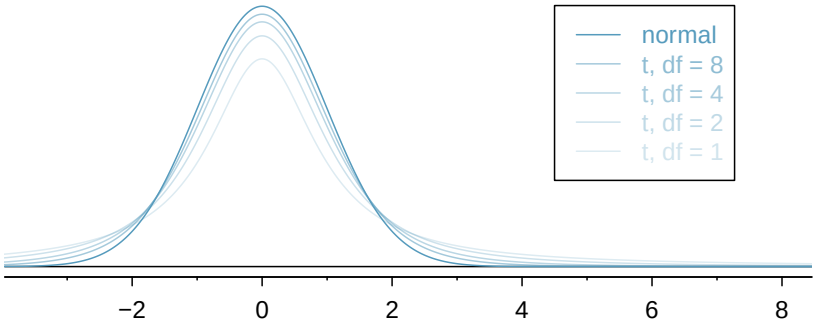


Figure 7.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal model.

When the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution. In Section 7.1.3, we relate degrees of freedom to sample size.

It's very useful to become familiar with the t-distribution, because it allows us greater flexibility than the normal distribution when analyzing numerical data. In practice, it's common to use statistical software, such as **R**, Python, or SAS. Alternatively, a graphing calculator or a **t-table** may be used; the t-table is similar to the normal distribution table, and it may be found in Appendix B.2 for those who wish to use this option. No matter the approach you choose, apply your method using the examples below to confirm your understanding of working with tail areas in the t-distribution.

EXAMPLE 7.1

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

(E)

Just like a normal probability problem, we first draw the picture in Figure 7.3 and shade the area below -2.10. If this were a normal distribution, the area would be a little less than 0.025, since about 5% of the area under a normal curve goes out beyond ± 1.96 standard deviations in the two tails. For a t-distribution, the tails are a little thicker, and using statistical software, we can obtain a precise value: 0.0250. The tail area below -2.10 in the t-distribution with $df = 18$ is the same as the tail area below -1.96 in the normal distribution.

EXAMPLE 7.2

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure 7.4. Estimate the proportion of the distribution falling above 1.65.

(E)

With a normal distribution, this would correspond to about 0.05, so we should expect the t-distribution to give us a value in this neighborhood. Using statistical software, we can obtain the value as 0.0573.

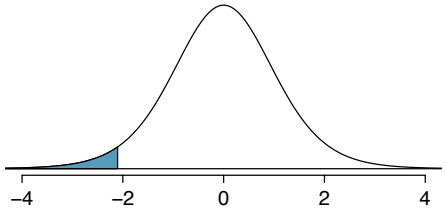


Figure 7.3: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

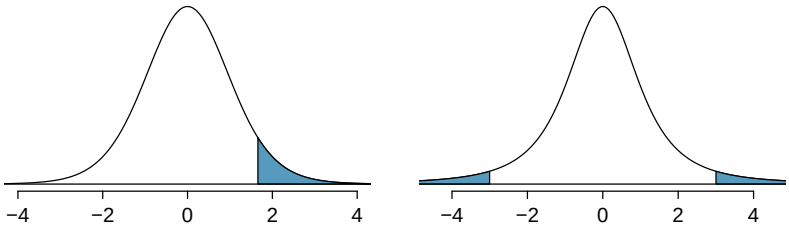


Figure 7.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

EXAMPLE 7.3

A t-distribution with 2 degrees of freedom is shown in the right panel of Figure 7.4. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

(E)

With so few degrees of freedom, the t-distribution will give a more notably different value than the normal distribution. Under a normal distribution, the area would be about 0.003 using the 68-95-99.7 rule. For a t-distribution with $df = 2$, the area in both tails beyond 3 units totals 0.0071. While the absolute amount is similar, this area is more than double that of from the normal distribution.

GUIDED PRACTICE 7.4

(G) What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units? Use your preferred method for finding tail areas.[2]

7.1.3 Conditions for using the t-distribution for inference on a sample mean

To proceed with the t-distribution for inference about a single mean, we first check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample, or if the data are from an experiment or random process, we check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t-distribution with $n - 1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t-distribution with $df = 19 - 1 = 18$ degrees of freedom and proceed in the same way as we did in Chapter 5, except that *now we use the t-distribution.*

[2] We find the shaded area *above* -1.79 (we leave the picture to you). The lower tail area, which is the part we are not trying to find, has an area of 0.0447, so the upper area would have an area of $1 - 0.0447 = 0.9553$.

7.1.4 One sample t-confidence intervals

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them.

Figure 7.5: A Risso's dolphin.

Photo by Mike Baird (www.bairdphotos.com). CC BY 2.0 license.

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan. The data are summarized in Figure 7.6. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Figure 7.6: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

EXAMPLE 7.5

Are the independence and normality conditions satisfied for this data set?

(E)

The observations are a simple random sample, therefore independence is reasonable. The summary statistics in Figure 7.6 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

In the normal model, we used $z^\star$ and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t-distribution:

$$\bar{x} \pm t^\star_{df} \times SE$$

The sample mean and estimated standard error are computed just as before ($\bar{x} = 4.4$ and $SE = s/\sqrt{n} = 0.528$). The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t-distribution with df degrees of freedom. Before determining this cutoff, we will first need the degrees of freedom.

DEGREES OF FREEDOM FOR A SINGLE SAMPLE

If the sample has n observations and we are examining a single mean, then we use the t-distribution with $df = n - 1$ degrees of freedom.

In our current example, we should use the t-distribution with $df = 19 - 1 = 18$ degrees of freedom. We can generally identify $t^\star_{18}$ using statistical software. Alternatively, we could use the t-table in Appendix B.2. Generally the value of $t^\star_{df}$ is slightly larger than what we would get under the normal model with $z^\star$.

Finally, we can substitute all our values into the confidence interval equation to create the 95% confidence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \pm t^\star_{18} \times SE \quad \rightarrow \quad 4.4 \pm 2.10 \times 0.528 \quad \rightarrow \quad (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram, which is considered extremely high.

FINDING A T-CONFIDENCE INTERVAL FOR THE MEAN

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \pm t^\star_{df} \times SE$$

where $\bar{x}$ is the sample mean, $t^\star_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

GUIDED PRACTICE 7.6

(G) The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[3]

EXAMPLE 7.7

(E) Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Guided Practice 7.6. If we are to use the t-distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^\star_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: $df = n - 1 = 14$.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row $df = 14$, we identify $t^\star_{14} = 1.76$.

[3] There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red flags for the normal model based on this (limited) information, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

CONFIDENCE INTERVAL FOR A SINGLE MEAN

Once you've determined a one-mean confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\bar{x}$, s, n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, find $t^\star_{df}$, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

GUIDED PRACTICE 7.8

Using the results of Guided Practice 7.6 and Example 7.7, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[4]

7.1.5 One sample t-tests

Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Race, which is a 10-mile race in Washington, DC each spring.

The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine using data from 100 participants in the 2017 Cherry Blossom Race whether runners in this race are getting faster or slower, versus the other possibility that there has been no change.

GUIDED PRACTICE 7.9

What are appropriate hypotheses for this context?[5]

GUIDED PRACTICE 7.10

The data come from a simple random sample of all participants, so the observations are independent. However, should we be worried about skew in the data? See Figure 7.7 for a histogram of the differences.[6]

With independence satisfied and slight skew not a concern for this large of a sample, we can proceed with performing a hypothesis test using the t-distribution.

GUIDED PRACTICE 7.11

The sample mean and sample standard deviation of the sample of 100 runners from the 2017 Cherry Blossom Race are 97.32 and 16.98 minutes, respectively. Recall that the sample size is 100. What is the p-value for the test, and what is your conclusion?[7]

[4] $\bar{x} \pm t^\star_{14} \times SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

[5] H_0: The average 10 mile run time was the same for 2006 and 2017. $\mu = 93.29$ minutes. H_A: The average 10 mile run time for 2017 was *different* than that of 2006. $\mu \neq 93.29$ minutes.

[6] With a sample of 100, we should only be concerned if there is very strong skew. The histogram of the data doesn't show any notable skew.

[7] With the conditions satisfied for the t-distribution, we can compute the standard error ($SE = 16.98/\sqrt{100} = 1.70$ and the *T-score*: $T = \frac{97.32-93.29}{1.70} = 2.37$. (There is more on this after the guided practice, but a T-score and Z-score are calculated in the same way.) For $df = 100 - 1 = 99$, we would find $T = 2.37$ to falls right on the fourth column in the 90 degrees of freedom row, which corresponds to a two-tailed p-value of 0.02. The p-value could also have been calculated with statistical software and 99 degrees of freedom as 0.0195. Because the p-value is smaller than 0.05, we reject the null hypothesis. That is, the data provide strong evidence that the average run time for the Cherry Blossom Run in 2017 is different than the 2006 average. Since the observed value is above the null value and we have

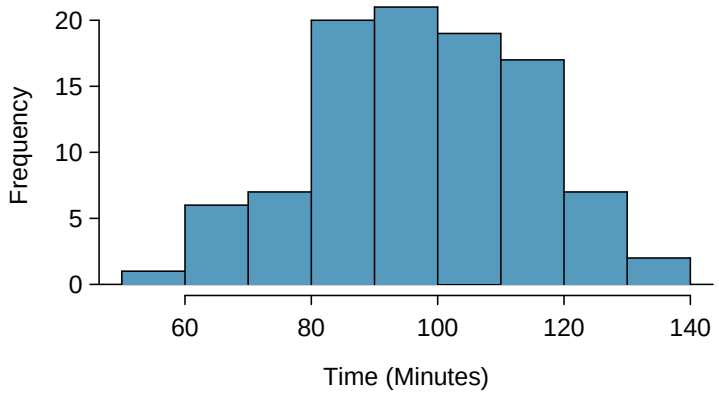


Figure 7.7: A histogram of `time` for the sample Cherry Blossom Race data.

To help us remember to use the t-distribution, we use a T to represent the test statistic, and we often call this a **T-score**. The Z-score and T-score are computed in the exact same way and are conceptually identical: each represents how many standard errors the observed value is from the null value.

HYPOTHESIS TESTING FOR A SINGLE MEAN

Once you've determined a one-mean hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list out hypotheses, identify the significance level, and identify $\bar{x}$, s, and n.

Check. Verify conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, compute the T-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

---

rejected the null hypothesis, we would conclude that runners in the race were slower on average in 2017 than in 2006.

7.2 Paired data

In earlier editions of this textbook, we found that Amazon prices were, on average, lower than those of the UCLA Bookstore for UCLA courses in 2010. It's been several years, and many stores have adapted to the online market, so we wondered, how is the UCLA Bookstore doing today?

We sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
67	Korean	1	24.96	23.79	1.17
68	Jewish Studies	M10	35.96	32.40	3.56

Figure 7.8: Five cases of the `textbooks` data set.

7.2.1 Paired observations

Each textbook has two corresponding prices in the data set: one for the UCLA Bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

PAIRED DATA

Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the `textbook` data set, we look at the differences in prices, which is represented as the `diff` variable in the `textbooks` data. Here the differences are taken as

$$\text{UCLA Bookstore price} - \text{Amazon price}$$

for each book. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure 7.9. Using differences between paired observations is a common and useful way to analyze paired data.

GUIDED PRACTICE 7.12

G The first difference shown in Figure 7.8 is computed as $47.97 - 47.45 = 0.52$. Verify the differences are calculated correctly for observations 2 and 3.[8]

7.2.2 Inference for paired data

To analyze a paired data set, we simply analyze the differences. We can use the same t-distribution techniques we applied in the last section.

[Consider breaking the next example into two pieces.]

[8] Observation 2: $14.26 - 13.55 = 0.71$. Observation 3: $13.50 - 12.53 = 0.97$.

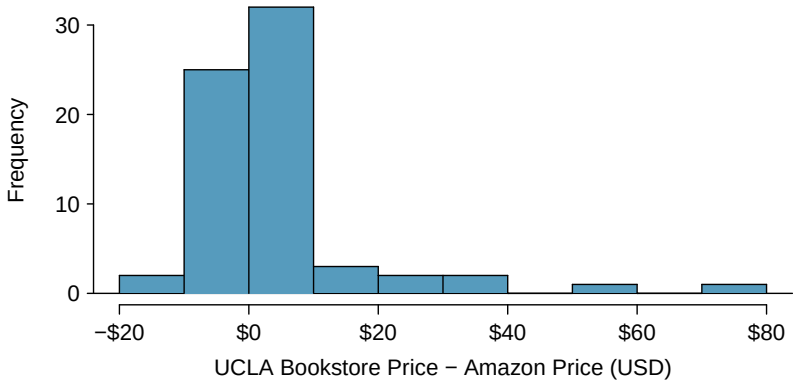


Figure 7.9: Histogram of the difference in price for each book sampled. These data are very strongly skewed.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

Figure 7.10: Summary statistics for the price differences. There were 68 books, so there are 68 differences.

EXAMPLE 7.13

Set up and implement a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price.

We are considering two scenarios: there is no difference or there is some difference in average prices.

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.
H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Can the t-distribution be used for this application? The observations are based on a simple random sample, so independence is reasonable. While the distribution is very strongly skewed, we do have $n = 68$ observations, and it's reasonable to move forward here. Because the conditions are reasonably satisfied, we can apply the t-distribution to this setting.

(E) We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 13.42$) and the number of differences ($n_{diff} = 68$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, which is shown in Figure 7.11. The p-value is represented by the two shaded tails.

To find the tail areas, we compute the test statistic, which is the T-score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$T = \frac{\bar{x}_{diff} - 0}{SE_{x_{diff}}} = \frac{3.58 - 0}{1.63} = 2.20$$

The degrees of freedom are $df = 68 - 1 = 67$. Using statistical software, we would find the one-tail area is 0.0156. Doubling this area gives the p-value: 0.0312. Because the p-value is less than 0.05, we reject the null hypothesis. Amazon prices are, on average, lower than the UCLA Bookstore prices for UCLA courses.

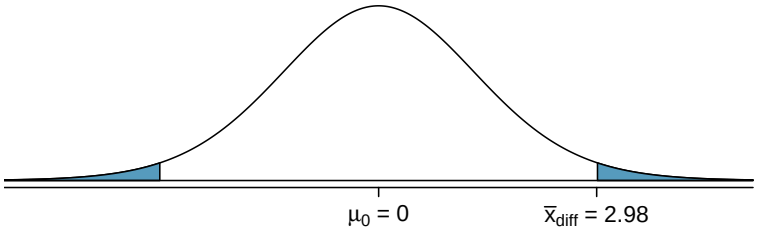


Figure 7.11: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

GUIDED PRACTICE 7.14

G Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[9]

GUIDED PRACTICE 7.15

G We have strong evidence to conclude Amazon is, on average, less expensive. How should this conclusion affect UCLA students' buying habits? Should UCLA students always buy their books on Amazon?[10]

[9] Conditions have already verified and the standard error computed in Example 7.13. To find the interval, identify $t^\star_{67}$ using statistical software or the t-table ($t^\star_{67} = 2.00$), and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 3.58 \pm 2.00 \times 1.63 \quad \rightarrow \quad (0.326.84)$$

We are 95% confident that Amazon is, on average, between \$0.32 and \$6.84 less expensive than the UCLA Bookstore for UCLA course books.

[10] The average price difference is only mildly useful for this question. Examine the distribution shown in Figure 7.9. There are certainly a handful of cases where Amazon prices are much below the UCLA Bookstore's, which suggests it is worth checking Amazon or other online sites before purchasing. However, in many cases the Amazon price is above what the UCLA Bookstore charges, and most of the time the price isn't that different. Ultimately, if getting a book immediately from the bookstore is notably more convenient, e.g. to get started on homework, it's likely a good idea to go with the UCLA Bookstore unless the price difference on a specific book happens to be quite large.

For reference, this is a very different result from what we (the authors) had seen in a similar data set from 2010. At that time, Amazon prices were almost uniformly lower than those of the UCLA Bookstore's and by a large margin, making the case to use Amazon over the UCLA Bookstore quite compelling at that time.

7.3 Difference of two means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. Just as with a single sample, we identify conditions to ensure we can use the t-distribution with a point estimate of the difference, $\bar{x}_1 - \bar{x}_2$.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the impact of pregnant womens' smoking habits on birth weights of newborns, and exploring whether there is statistically significant evidence that one variation of an exam is harder than another variation. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

7.3.1 Confidence interval for a difference of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Figure 7.12 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

A point estimate of the difference in the heart pumping variable can be found using the difference in the sample means:

$$\bar{x}_{esc} - \bar{x}_{control} = 3.50 - (-4.33) = 7.83$$

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Figure 7.12: Summary statistics of the embryonic stem cell study.

USING THE T-DISTRIBUTION FOR A DIFFERENCE IN MEANS

The t-distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t-distribution and (2) the samples are independent.

EXAMPLE 7.16

Can the t-distribution be used to make inference using the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$?

We check the two required conditions:

1. In this study, the sheep were independent of each other. Additionally, the distributions in Figure 7.13 don't show any clear deviations from normality, where we watch for prominent outliers in particular for such small samples. These findings imply each sample mean could itself be modeled using a t-distribution.

2. The sheep in each group were also independent of each other.

Because both conditions are met, we can use the t-distribution to model the difference of the two sample means.

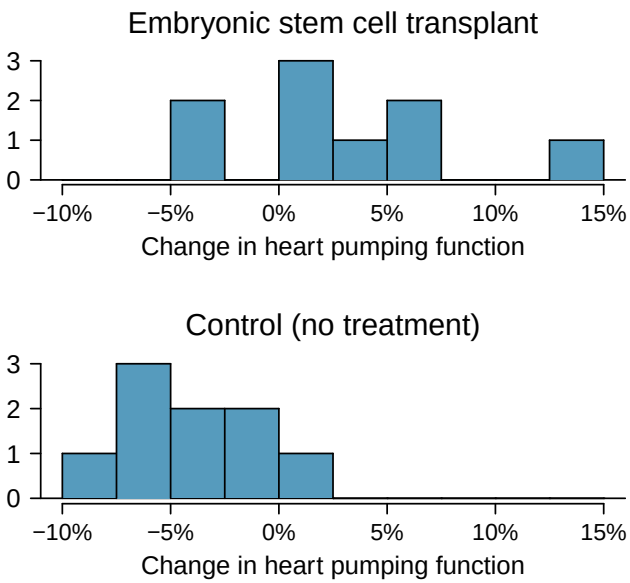


Figure 7.13: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

We can quantify the variability in the point estimate, $\bar{x}_{esc} - \bar{x}_{control}$, using the following formula for its standard error:

$$SE_{\bar{x}_{esc}-\bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$

We usually estimate this standard error using standard deviation estimates based on the samples:

$$\begin{aligned} SE_{\bar{x}_{esc}-\bar{x}_{control}} &= \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}} \\ &\approx \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95 \end{aligned}$$

Because we will use the t-distribution, we also must identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will typically apply in the examples and guided practice.[11]

DISTRIBUTION OF A DIFFERENCE OF SAMPLE MEANS

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t-distribution and the standard error

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

when each sample mean can itself be modeled using a t-distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

[11] This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method. In this example, computer software would have provided us a more precise degrees of freedom of $df = 12.225$.

EXAMPLE 7.17

Calculate a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep after they've suffered a heart attack.

We will use the sample difference and the standard error for that point estimate from our earlier calculations:

(E)

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83$$
$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Using $df = 8$, we can identify the appropriate $t^\star_{df} = t^\star_8$ for a 95% confidence interval as 2.31. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \ \pm \ t^\star \times SE \quad \rightarrow \quad 7.83 \ \pm \ 2.31 \times 1.95 \quad \rightarrow \quad (3.32, 12.34)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.32% to 12.34%.

As with past statistical inference applications, there is a well-trodden procedure.

1. Prepare: retrieve critical contextual information, and if appropriate, set up hypotheses.
2. Check: ensure the required conditions are reasonably satisfied.
3. Calculate: find the standard error, and then construct a confidence interval or find a test statistic and p-value.
4. Conclude: interpret the result in the context of the application.

The details change a little from one setting to the next, but this general approach remain the same.

7.3.2 Hypothesis tests based on a difference in means

A data set called `ncbirths` represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Figure 7.14. We are particularly interested in two variables: `weight` and `smoke`. The `weight` variable represents the weights of the newborns and the `smoke` variable describes which mothers smoked during pregnancy. We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 7.15.

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Figure 7.14: Four cases from the `ncbirths` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

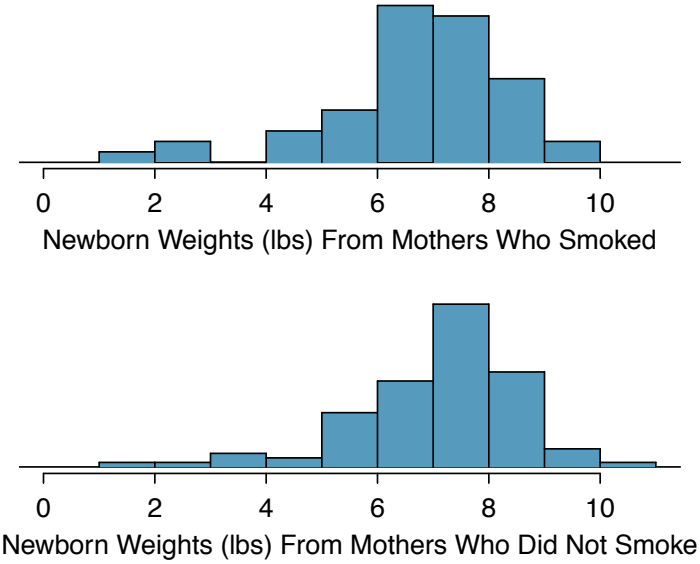


Figure 7.15: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

EXAMPLE 7.18

Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

The null hypothesis represents the case of no difference between the groups.

(E)

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.

H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

We check the two conditions necessary to apply the t-distribution to the difference in sample means. (1) Because the data come from a simple random sample, the observations are independent. Additionally, while each distribution is strongly skewed, the sample sizes of 50 and 100 would make it reasonable to model each mean separately using a t-distribution. The skew is reasonable for these sample sizes of 50 and 100. (2) The independence reasoning applied in (1) also ensures the observations in each sample are independent. Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

	`smoker`	`nonsmoker`
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Figure 7.16: Summary statistics for the `ncbirths` data set.

GUIDED PRACTICE 7.19

The summary statistics in Figure 7.16 may be useful for this exercise.

(G)

(a) What is the point estimate of the population difference, $\mu_n - \mu_s$?

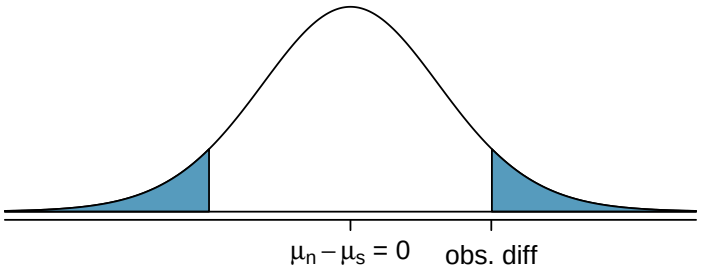
(b) Compute the standard error of the point estimate from part (a).[12]

EXAMPLE 7.20

Draw a picture to represent the p-value for the hypothesis test from Example 7.18.

To depict the p-value, we draw the distribution of the point estimate as though H_0 were true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.

E



EXAMPLE 7.21

Compute the p-value of the hypothesis test using the figure in Example 7.20, and evaluate the hypotheses using a significance level of $\alpha = 0.05$.

We start by computing the T-score:

E

$$T = \frac{0.40 - 0}{0.26} = 1.54$$

Next, we find the single tail area using software (or the table in Appendix B.2 on page 287). We'll use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The one tail area is 0.065; doubling this value gives the two-tail area and p-value, 0.135. This p-value is larger than the significance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

GUIDED PRACTICE 7.22

G We've seen much research suggesting smoking is harmful during pregnancy, so how could we fail to reject the null hypothesis in Example 7.21? [13]

GUIDED PRACTICE 7.23

G If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect the difference?[14]

Public service announcement: while we have used this relatively small data set as an example, larger data sets show that women who smoke tend to have smaller newborns. In fact, some in the tobacco industry actually had the audacity to tout that as a *benefit* of smoking:

[12] (a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$. (b) The standard error of the estimate can be estimated using the standard error formula:

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$

[13] It is possible that there is some difference but we did not detect it. If there is a difference, we made a Type 2 Error. Notice: we also don't have enough information to, if there is an actual difference, confidently say which direction that difference would be in.

[14] We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists. In fact, this is exactly what we would find if we examined a larger data set!

> *It's true. The babies born from women who smoke are smaller, but they're just as healthy as the babies born from women who do not smoke. And some women would prefer having smaller babies.*
>
> - Joseph Cullman, Philip Morris' Chairman of the Board
> on CBS' *Face the Nation*, Jan 3, 1971

Fact check: the babies from women who smoke are not actually as healthy as the babies from women who do not smoke.[15]

7.3.3 Case study: two versions of a course exam

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, she shuffled the exams together to ensure each student received a random version. Summary statistics for how students performed on these two exams are shown in Figure 7.17. Anticipating complaints from students who took Version B, she would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Figure 7.17: Summary statistics of scores for each exam version.

GUIDED PRACTICE 7.24

G Construct a hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance. We will later evaluate these hypotheses using $\alpha = 0.01$.[16]

GUIDED PRACTICE 7.25

To evaluate the hypotheses in Guided Practice 7.24 using the t-distribution, we must first verify assumptions.[17]

G (a) Does it seem reasonable that the scores are independent within each group?

(b) What about the normality / skew condition for observations in each group?

(c) Do you think scores from the two groups would be independent of each other, i.e. the two samples are independent?

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t-distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated as

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62$$

[15] You can watch an episode of John Oliver on *Last Week Tonight* to explore the present day offenses of the tobacco industry. Please be aware that there is some adult language: youtu.be/6UsHHOCH4q8.

[16] Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. H_A: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

[17] (a) It is probably reasonable to conclude the scores are independent, provided there was no cheating. (b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable. (c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1$ and $n_2 - 1$: $df = 26$.

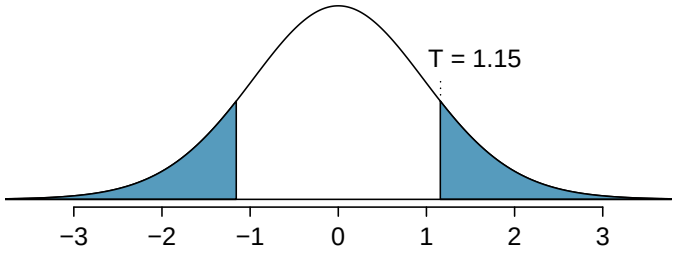


Figure 7.18: The t-distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

EXAMPLE 7.26

Identify the p-value using $df = 26$ and provide a conclusion in the context of the case study.

(E)

Using software, we can find the one-tail area (0.13) and then double this value to get the two-tail area, which is the p-value: 0.26. (Alternatively, we could use the t-table in Appendix B.2.) We examine row $df = 26$ in the t-table. In Guided Practice **??**, we specified that we would use $\alpha = 0.01$. Since the p-value is larger than α, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

7.3.4 Pooled standard deviation estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make the t-distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s^2_{pooled} = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t-distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$, if the standard deviations of the two groups are equal.

POOL STANDARD DEVIATIONS ONLY AFTER CAREFUL CONSIDERATION

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

7.4 Power calculations for a difference of means

Often times in experiment planning, there are two competing considerations:

- We want to collect enough data that we can detect important effects.
- Collecting data can be expensive, and in experiments involving people, there may be some risk to patients.

In this section, we focus on the context of a clinical trial, which is a health-related experiment where the subject are people, and we will determine an appropriate sample size where we can be 80% sure that we would detect any practically important effects.[18]

7.4.1 Going through the motions of a test

We're going to go through the motions of a hypothesis test. This will help us frame our calculations for determining an appropriate sample size for the study.

EXAMPLE 7.27

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial (experiment) to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication. People in the control group will continue to take their current medication through generic-looking pills to ensure blinding. Write down the hypotheses for a two-sided hypothesis test in this context.

Generally, clinical trials use a two-sided alternative hypothesis, so below are suitable hypotheses for this context:

(E)

H_0: The new drug performs exactly as well as the standard medication.
$\mu_{trmt} - \mu_{ctrl} = 0$.

H_A: The new drug's performance differs from the standard medication.
$\mu_{trmt} - \mu_{ctrl} \neq 0$.

Some researchers might argue for a one-sided test here, where the alternative would consider only whether the new drug performs better than the standard medication. However, it would be very informative to know whether the new drug performs worse than the standard medication, so we use a two-sided test to consider this possibility during the analysis.

[18] Even though we don't cover it explicitly, similar sample size planning is also helpful for observational studies.

EXAMPLE 7.28

The researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric.[a] If we had 100 patients per group, what would be the approximate standard error for $\bar{x}_{trmt} - \bar{x}_{ctrl}$?

The standard error is calculated as follows:

(E)

$$SE_{\bar{x}_{trmt}-\bar{x}_{ctrl}} = \sqrt{\frac{s_{trmt}^2}{n_{trmt}} + \frac{s_{ctrl}^2}{n_{ctrl}}} = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

This may be an imperfect estimate of $SE_{\bar{x}_{trmt}-\bar{x}_{ctrl}}$, since the standard deviation estimate we used may not be correct for this group of patients. However, it is sufficient for our purposes.

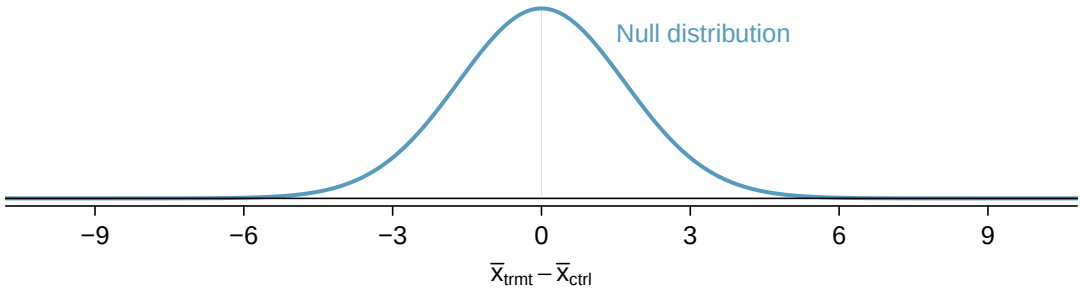
[a] In this particular study, we'd generally measure each patient's blood pressure at the beginning and end of the study, and then the outcome measurement for the study would be the average change in blood pressure. That is, both μ_{trmt} and μ_{ctrl} would represent average differences. This is what you might think of as a 2-sample paired testing structure, and we'd analyze it exactly just like a hypothesis test for a difference in the average change for patients. In the calculations we perform here, we'll suppose that 12 mmHg is the predicted standard deviation of a patient's blood pressure difference over the course of the study.

EXAMPLE 7.29

What does the null distribution of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ look like?

The degrees of freedom are greater than 30, so the distribution of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ will be approximately normal. The standard deviation of this distribution (the standard error) would be about 1.70, and under the null hypothesis, its mean would be 0.

(E)



EXAMPLE 7.30

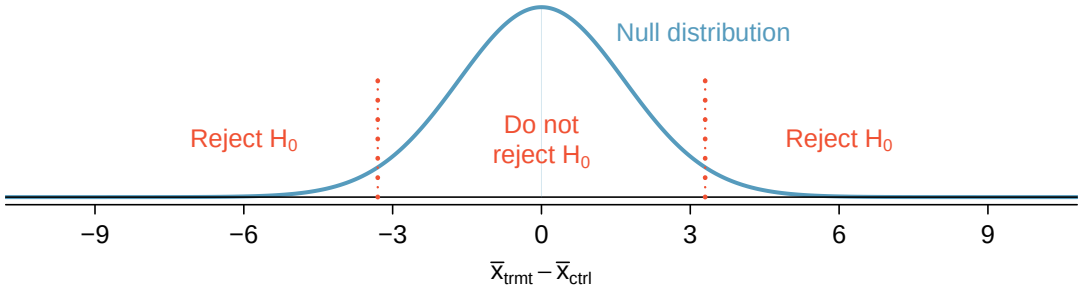
For what values of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ would we reject the null hypothesis?

For $\alpha = 0.05$, we would reject H_0 if the difference is in the lower 2.5% or upper 2.5% tail:

Lower 2.5%: For the normal model, this is 1.96 standard errors below 0, so any difference smaller than $-1.96 \times 1.70 = -3.332$ mmHg.

Upper 2.5%: For the normal model, this is 1.96 standard errors above 0, so any difference larger than $1.96 \times 1.70 = 3.332$ mmHg.

(E) The boundaries of these **rejection regions** are shown below:



Next, we'll perform some hypothetical calculations to determine the probability we reject the null hypothesis, if the alternative hypothesis were actually true.

7.4.2 Computing the power for a 2-sample test

When planning a study, we want to know how likely we are to detect an effect we care about. In other words, if there is a real effect, and that effect is large enough that it has practical value, then what's the probability that we detect that effect? This probability is called the **power**, and we can compute it for different sample sizes or for different *effect sizes*.

We first determine what is a practically significant result. Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. Here, 3 mmHg is the minimum **effect size** of interest, and we want to know how likely we are to detect this size of an effect in the study.

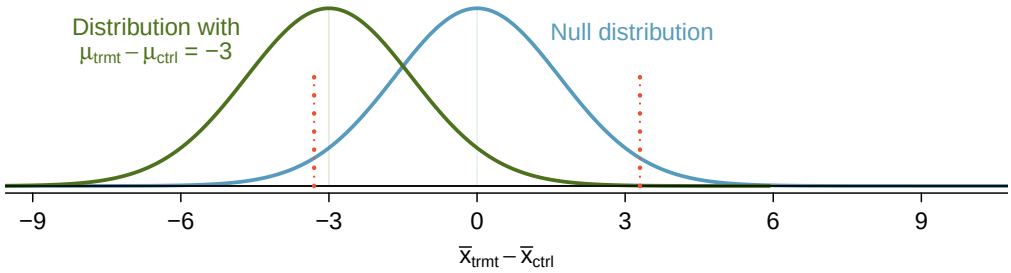
EXAMPLE 7.31

Suppose we decided to move forward with 100 patients per treatment group and the new drug reduces blood pressure by an additional 3 mmHg relative to the standard medication. What is the probability that we detect a drop?

Before we even do any calculations, notice that if $\bar{x}_{trmt} - \bar{x}_{ctrl} = -3$ mmHg, there wouldn't even be sufficient evidence to reject H_0. That's not a good sign.

To calculate the probability that we will reject H_0, we need to determine a few things:

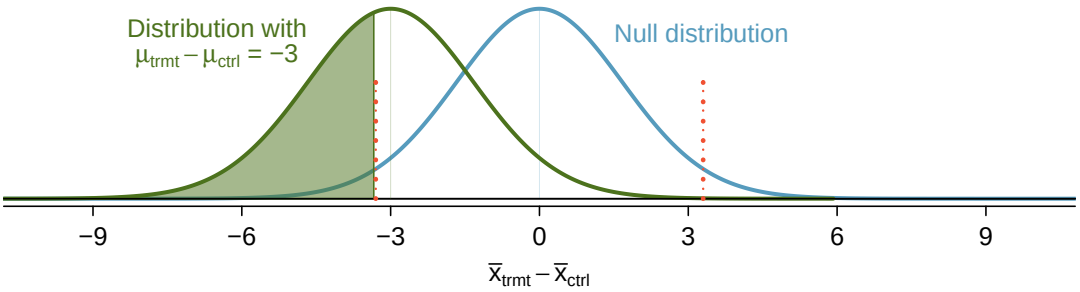
- The sampling distribution for $\bar{x}_{trmt} - \bar{x}_{ctrl}$ when the true difference is -3 mmHg. This is the same as the null distribution, except it is shifted to the left by 3:



(E)

- The rejection regions, which are outside of the dotted lines above.
- The fraction of the distribution that falls in the rejection region.

In short, we need to calculate the probability that $x < -3.332$ for a normal distribution with mean -3 and standard deviation 1.7. To do so, we first shade the area we want to calculate:



Then we calculate the T-score and find the tail area using either statistical software or the probability table with $df = 99$ (from the minimum of $n_1 - 1$ and $n_2 - 1$):

$$T = \frac{-3.332 - (-3)}{1.7} = -0.20 \qquad \rightarrow \qquad 0.4207$$

The power for the test is about 42% when $\mu_{trmt} - \mu_{ctrl} = -3$ and each group has a sample size of 100.

In Example 7.31, we ignored the upper rejection region in the calculation, which was in the opposite direction of the hypothetical truth, i.e. -3. The reasoning? There wouldn't be any value in rejecting the null hypothesis and concluding there was an increase when in fact there was a decrease.

7.4.3 Determining a proper sample size

In the last example, we found that if we have a sample size of 100 in each group, we can only detect an effect size of 3 mmHg with a probability of about 0.42. Suppose the researchers moved forward and only used 100 patients per group, and the data did not support the alternative

hypothesis, i.e. the researchers did not reject H_0. This is a very bad situation to be in for a few reasons:

- In the back of the researchers' minds, they'd all be wondering, *maybe there is a real and meaningful difference, but we weren't able to detect it with such a small sample.*
- The company probably invested hundreds of millions of dollars in developing the new drug, so now they are left with great uncertainty about its potential since the experiment didn't have a great shot at detecting effects that could still be important.
- Patients were subjected to the drug, and we can't even say with much certainty that the drug doesn't help (or harm) patients.
- Another clinical trial may need to be run to get a more conclusive answer as to whether the drug does hold any practical value, and conducting a second clinical trial may take years and many millions of dollars.

We want to avoid this situation, so we need to determine an appropriate sample size to ensure we can be pretty confident that we'll detect any effects that are practically important. As mentioned earlier, a change of 3 mmHg was deemed to be the minimum difference that was practically important. As a first step, we could calculate power for several different sample sizes. For instance, let's try 500 patients per group.

GUIDED PRACTICE 7.32

Calculate the power to detect a change of -3 mmHg when using a sample size of 500 per group.[19]

(a) Determine the standard error (recall that the standard deviation for patients was expected to be about 12 mmHg).

(b) Identify the null distribution and rejection regions.

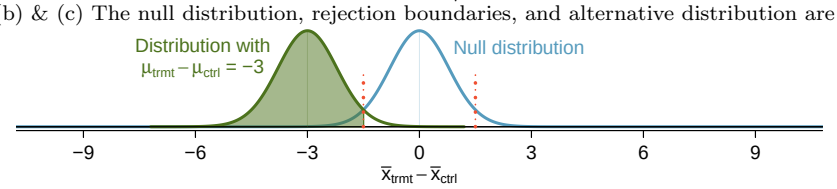
(c) Identify the alternative distribution when $\mu_{trmt} - \mu_{ctrl} = -3$.

(d) Compute the probability we reject the null hypothesis.

The researchers decided 3 mmHg was the minimum difference that was practically important, and with a sample size of 500, we can be very certain (97.7% or better) that we will detect any such difference. We now have moved to another extreme where we are exposing an unnecessary number of patients to the new drug in the clinical trial. Not only is this ethically questionable, but it would also cost a lot more money than is necessary to be quite sure we'd detect any important effects.

The most common practice is to identify the sample size where the power is around 80%, and sometimes 90%. Other values may be reasonable for a specific context, but 80% and 90% are most commonly targeted as a good balance between high power and not exposing too many patients to a new treatment (or wasting too much money). We could compute the power of the test at several other possible sample sizes until we find one that's close to 80%, but that's inefficient. Instead, we should solve the problem backwards.

[19] (a) The standard error is given as $SE = \sqrt{\frac{12^2}{500} + \frac{12^2}{500}} = 0.76$.
(b) & (c) The null distribution, rejection boundaries, and alternative distribution are shown below:

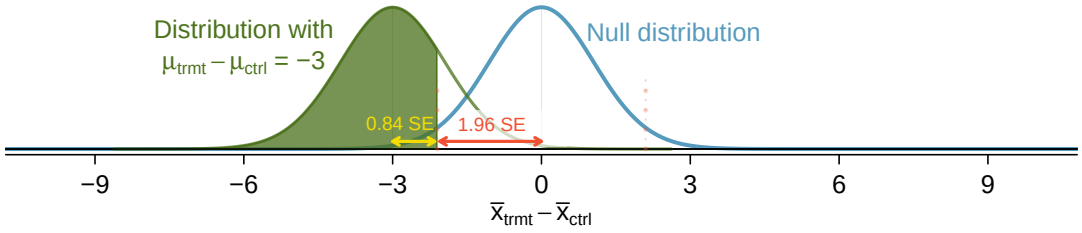


The rejection regions are the areas on the outside of the two dotted lines and are at $\pm 0.76 \times 1.96 = \pm 1.49$.
(d) The area of the alternative distribution where $\mu_{trmt} - \mu_{ctrl} = -3$ has been shaded. We compute the Z-score and find the tail area: $T = \frac{-1.49-(-3)}{0.76} = 1.99 \rightarrow 0.9767$ (can use $df = 500$ from the minimum of the two sample sizes minus 1), which is the power of the test for a difference of 3 mmHg. With 500 patients per group, we would be about 97.7% sure (or more) that we'd detect any effects that are at least 3 mmHg in size.

EXAMPLE 7.33

What sample size will lead to a power of 80%?

We start by identifying the T-score that would give us a lower tail of 80%. For a moderately large sample size per group, the T-score for a lower tail of 80% would be about $T = 0.84$. (If our calculations suggest a very sample size, we should recalculate this part and basically do the problem one more time.)



Additionally, the rejection region always extends $1.96 \times SE$ from the center of the null distribution for $\alpha = 0.05$ (again supposing a larger sample size to get the value of 1.96). This allows us to calculate the target distance between the center of the null and alternative distributions in terms of the standard error:

(E)

$$0.84 \times SE + 1.96 \times SE = 2.8 \times SE$$

In our example, we also want the distance between the null and alternative distributions' centers to equal the minimum effect size of interest, 3 mmHg, which allows us to set up an equation between this difference and the standard error:

$$3 = 2.8 \times SE$$

$$3 = 2.8 \times \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = \frac{2.8^2}{3^2} \times \left(12^2 + 12^2\right) = 250.88$$

We should target about 251 patients per group.

Had the suggested sample size been relatively small – roughly 30 or smaller – it would have been a good idea to rework the calculations using the degrees of fredom for the smaller sample size under that initial sample size. That is, we would have revised the 0.84 and 1.96 values based on degrees of freedom implied by the initial sample size. The revised sample size target will generally be a little larger.

The standard error difference of $2.8 \times SE$ is specific to a context where the targeted power is 80% and the significance level is $\alpha = 0.05$. If the targeted power is 90% or if we use a different significance level, then we'll use something a little different than $2.8 \times SE$.

GUIDED PRACTICE 7.34

(G)

Suppose the targeted power was 90% and we were using $\alpha = 0.01$. How many standard errors should separate the centers of the null and alternative distribution, where the alternative distribution is centered at the minimum effect size of interest? Assume the test is two-sided.[20]

[20] First, find the T-score such that 90% of the distribution is below it: $Z = 1.28$ (assume large df). Next, find the cutoffs for the rejection regions: ± 2.58. Then the difference in centers should be about $1.28 \times SE + 2.58 \times SE = 3.86 \times SE$.

GUIDED PRACTICE 7.35

List some considerations that are important in determining what the power should be for an experiment.[21]

Figure 7.19 shows the power for sample sizes from 20 patients to 5,000 patients when $\alpha = 0.05$ and the true difference is -3. This curve was constructed by writing a program to compute the power for many different sample sizes.

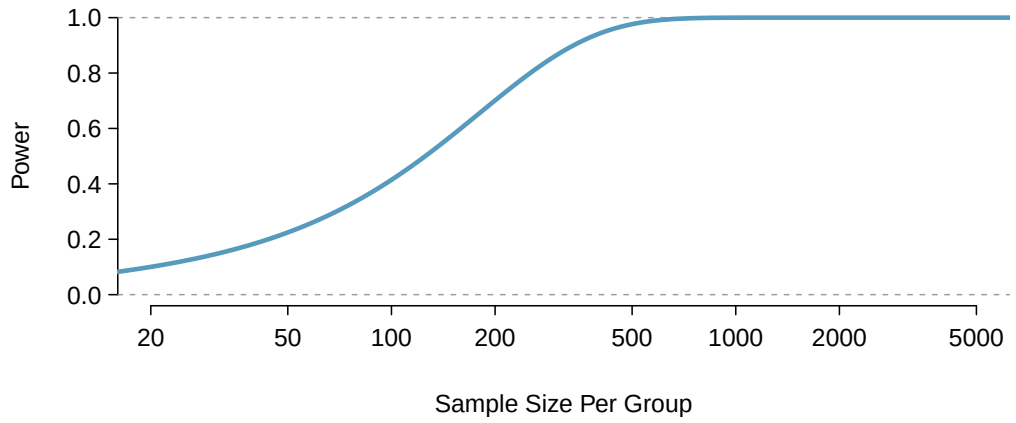


Figure 7.19: The curve shows the power for different sample sizes in the context of the blood pressure example when the true difference is -3. Having more than about 250 to 350 observations doesn't provide much additional value in detecting an effect when $\alpha = 0.05$.

Power calculations for expensive or risky experiments are critical. However, what about experiments that are inexpensive and where the ethical considerations are minimal? For example, if we are doing final testing on a new feature on a popular website, how would our sample size considerations change? As before, we'd want to make sure the sample is big enough. However, suppose the feature has undergone some testing and is known to perform well (e.g. the website's users seem to enjoy the feature). Then it may be reasonable to run a larger experiment if there's value from having a more precise estimate of the feature's effect, such as helping guide the development of the next useful feature.

[21] Answers will vary, but here are a few important considerations:

- Whether there is any risk to patients in the study.
- The cost of enrolling more patients.
- The potential downside of not detecting an effect of interest.

7.5 Comparing many means with ANOVA

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

EXAMPLE 7.36

College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

(E)

The hypotheses may be written in the following form:

H_0: The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.

H_A: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

EXAMPLE 7.37

Examine Figure 7.20. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

(E)

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large *relative to the variability in the individual observations within each group.*

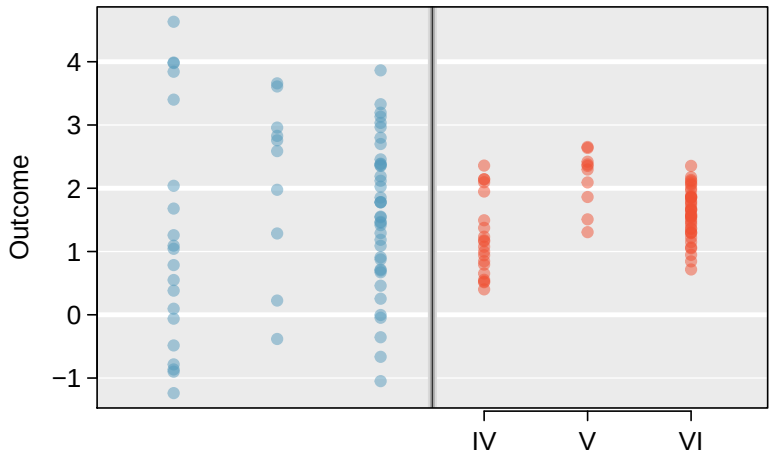


Figure 7.20: Side-by-side dot plot for the outcomes for six groups.

7.5.1 Is batting performance related to player position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: outfielder (`OF`), infielder (`IF`), and catcher (`C`). We will use a data set called `bat18`, which includes batting records of 429 Major League Baseball (MLB) players from the 2018 season who had at least 100 at bats. Six of the 429 cases represented in `bat18` are shown in Figure 7.21, and descriptions for each variable are provided in Figure 7.22. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (`OBP`). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	Abreu, J	CWS	IF	499	132	22	78	0.265	0.325
2	Acuna Jr., R	ATL	OF	433	127	26	64	0.293	0.366
3	Adames, W	TB	IF	288	80	10	34	0.278	0.348
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
427	Zimmerman, R	WSH	IF	288	76	13	51	0.264	0.337
428	Zobrist, B	CHC	IF	455	139	9	58	0.305	0.378
429	Zunino, M	SEA	C	373	75	20	44	0.201	0.259

Figure 7.21: Six cases from the `bat18` data matrix.

variable	description
`name`	Player name
`team`	The abbreviated name of the player's team
`position`	The player's primary field position (`OF`, `IF`, `C`)
`AB`	Number of opportunities at bat
`H`	Number of hits
`HR`	Number of home runs
`RBI`	Number of runs batted in
`AVG`	Batting average, which is equal to `H/AB`
`OBP`	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Figure 7.22: Variables and their descriptions for the `bat18` data set.

GUIDED PRACTICE 7.38

The null hypothesis under consideration is the following: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{C}}$. Write the null and corresponding alternative hypotheses in plain language.[22]

EXAMPLE 7.39

The player positions have been divided into four groups: outfield (`OF`), infield (`IF`), and catcher (`C`). What would be an appropriate point estimate of the on-base percentage by outfielders, $\mu_{\texttt{OF}}$?

A good estimate of the on-base percentage by outfielders would be the sample average of `OBP` for just those players whose position is outfield: $\bar{x}_{OF} = 0.320$.

Figure 7.23 provides summary statistics for each group. A side-by-side box plot for the on-base percentage is shown in Figure 7.24. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

	OF	IF	C
Sample size (n_i)	160	205	64
Sample mean ($\bar{x}_i$)	0.320	0.318	0.302
Sample SD (s_i)	0.043	0.038	0.038

Figure 7.23: Summary statistics of on-base percentage, split by player position.

EXAMPLE 7.40

The largest difference between the sample means is between the designated hitter and the outfielder positions. Consider again the original hypotheses:

H_0: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{C}}$

H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of $\mu_{\texttt{C}}$ and $\mu_{\texttt{OF}}$ is statistically significant at a 0.05 significance level?

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an inflation in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional information on the ideas expressed in Example 7.40, we recommend reading about the **prosecutor's fallacy**.[23]

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in sample means could have happened just by chance even if there was no difference in the respective population means.

[22] H_0: The average on-base percentage is equal across the four positions. H_A: The average on-base percentage varies across some (or all) groups.

[23] See, for example, andrewgelman.com/2007/05/18/the_prosecutors.

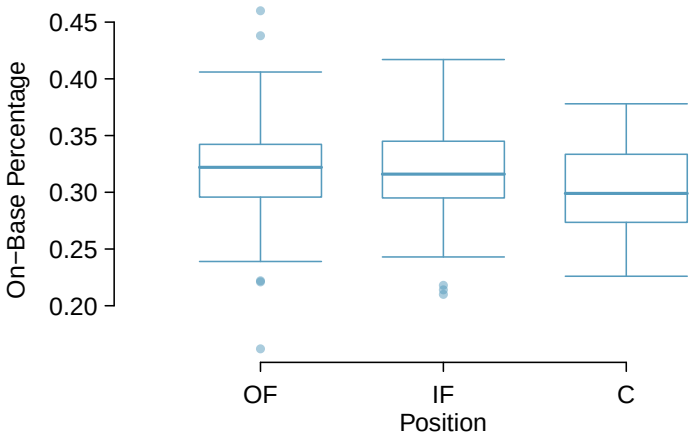


Figure 7.24: Side-by-side box plot of the on-base percentage for 429 players across four groups. There is one prominent outlier visible in the infield group, but with 154 observations in the infield group, this outlier is not a concern.

7.5.2 Analysis of variance (ANOVA) and the F test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups** (MSG), and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote,[24] however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error** (MSE), which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE are provided in the footnote[25] for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE}$$

[24] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i.

[25] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total** (SST) is computed as

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where the sum is over all observations in the data set. Then we compute the **sum of squared errors** (SSE) in one of two equivalent ways:

$$\begin{aligned} SSE &= SST - SSG \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \end{aligned}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

GUIDED PRACTICE 7.41

For the baseball data, $MSG = 0.00803$ and $MSE = 0.00158$. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 5.077.[26]

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 2 and 426 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 7.25.

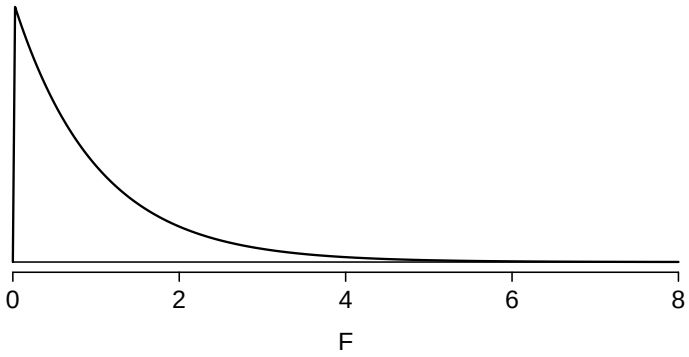


Figure 7.25: An F distribution with $df_1 = 3$ and $df_2 = 323$.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

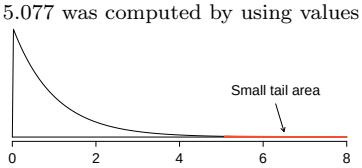
THE F STATISTIC AND THE F TEST

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

GUIDED PRACTICE 7.42

The test statistic for the baseball example is $F = 5.077$. Shade the area corresponding to the p-value in Figure 7.25. [27]

[26] There are $k = 3$ groups, so $df_G = k - 1 = 2$. There are $n = n_1 + n_2 + n_3 = 429$ total observations, so $df_E = n - k = 426$. Then the F statistic is computed as the ratio of MSG and MSE: $F = \frac{MSG}{MSE} = \frac{0.00803}{0.00158} = 5.082 \approx 5.077$. ($F = 5.077$ was computed by using values for MSG and MSE that were not rounded.)

[27]


EXAMPLE 7.43

The p-value corresponding to the shaded area in the solution of Guided Practice 7.42 is equal to about 0.0066. Does this provide strong evidence against the null hypothesis?

(E)

The p-value is smaller than 0.05, indicating the evidence is strong enough to reject the null hypothesis at a significance level of 0.05. That is, the data provide strong evidence that the average on-base percentage varies by player's primary field position.

7.5.3 Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we will see in Chapters 8 and 9. Figure 7.26 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F test statistic and p-value can be retrieved from the last columns.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	2	0.0161	0.0080	5.0766	0.0066
Residuals	426	0.6740	0.0016		
				$s_{pooled} = 0.040$ on $df = 423$	

Figure 7.26: ANOVA summary for testing whether the average on-base percentage differs across player positions.

7.5.4 Graphical diagnostics for an ANOVA analysis

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

Independence. If the data are a simple random sample, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). For example, in the MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations.

Approximately normal. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small when it is ironically difficult to check for non-normality. A histogram of the observations from each group is shown in Figure 7.27. Since each of the groups we're considering have relatively large sample sizes, what we're looking for are major outliers. None are apparent, so this conditions is reasonably met.

Constant variance. The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a side-by-side box plot of the outcomes across the groups, as in Figure 7.24 on page 216. In this case, the variability is similar in the four groups but not identical. We see in Table 7.23 on page 215 that the standard deviation doesn't vary much from one group to the next.

DIAGNOSTICS FOR AN ANOVA ANALYSIS

Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

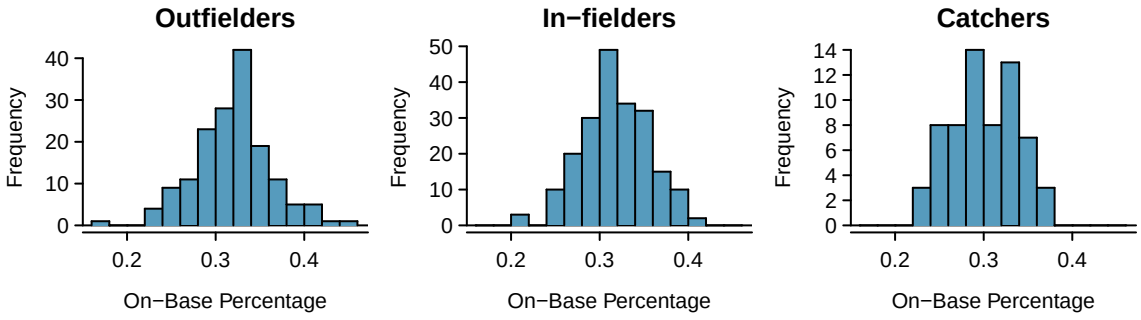


Figure 7.27: Histograms of OBP for each field position.

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.1	72.0	78.9
s_i	13.9	13.8	13.1

Figure 7.28: Summary statistics for the first midterm scores in three different lectures of the same course.

7.5.5 Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t-test, but we use a modified significance level and a pooled estimate of the standard deviation across groups. Usually this pooled standard deviation can be found in the ANOVA table, e.g. along the bottom of Figure 7.26.

EXAMPLE 7.44

Example 7.36 on page 213 discussed three statistics lectures, all taught during the same semester. Figure 7.28 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 7.29. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

(E) In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. No such situations were evident for these particular data, and we believe that independence is acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be verified in Figure 7.28, supporting the constant variance assumption.

GUIDED PRACTICE 7.45

(G) An ANOVA was conducted for the midterm data, and summary results are shown in Figure 7.30. What should we conclude?[28]

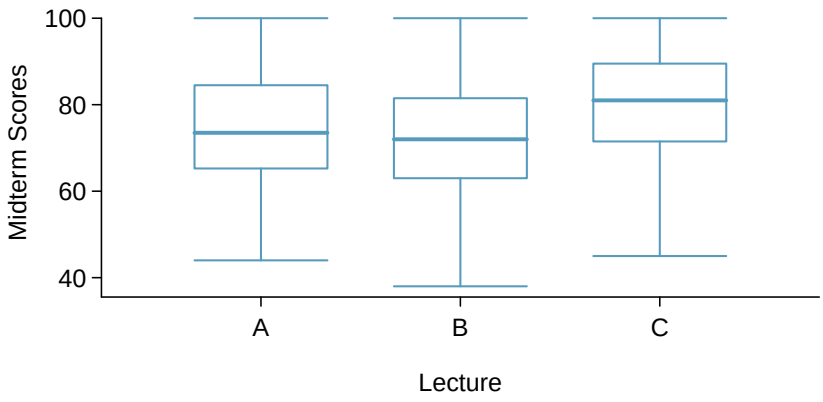


Figure 7.29: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		

$s_{pooled} = 13.61$ on $df = 161$

Figure 7.30: ANOVA summary table for the midterm data.

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t-test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 7.40 on page 215: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modified significance level.

MULTIPLE COMPARISONS AND THE BONFERRONI CORRECTION FOR α

The scenario of testing many pairs of groups is called **multiple comparisons**. The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha / K$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

[28] The p-value of the test is 0.0330, less than the default significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the difference in the average midterm scores are not due to chance.

EXAMPLE 7.46

In Guided Practice 7.45, you found strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modified significance level of $\alpha^* = 0.05/3 = 0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled} = 13.61$ on $df = 161$, which is provided in the ANOVA summary table.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{x}_A - \bar{x}_B = 75.1 - 72 = 3.1 \qquad SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56$$

(E)

(See Section 7.3.4 on page 204 for additional details.) This results in a T-score of 1.21 on $df = 161$ (we use the df associated with s_{pooled}). Statistical software was used to precisely identify the two-sided p-value since the modified significance level of 0.0167 is not found in the t-table. The p-value (0.228) is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on $df = 161$ and a two-sided p-value of 0.1462. This p-value is larger than α^*, so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on $df = 161$ and a two-sided p-value of 0.0102. This p-value is smaller than α^*. Here we find strong evidence of a difference in the means of lectures B and C.

We might summarize the findings of the analysis from Example 7.46 using the following notation:

$$\mu_A \stackrel{?}{=} \mu_B \qquad \mu_A \stackrel{?}{=} \mu_C \qquad \mu_B \neq \mu_C$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

REJECT H_0 WITH ANOVA BUT FIND NO GROUP DIFFERENCES

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion.* It only means we have not been able to successfully identify which specific groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

Chapter 8

Introduction to linear regression

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with straight lines are overlaid on scatterplots. Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

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

8.1 Fitting a line, residuals, and correlation

It's helpful to think deeply about the line fitting process. In this section, we define the form of a linear model, explore criteria for what makes a good fit, and introduce a new statistic called *correlation*.

8.1.1 Fitting a line to data

Figure 8.1 shows two variables whose relationship can be modeled perfectly with a straight line. The equation for the line is

$$y = 5 + 64.96x$$

Consider what a perfect linear relationship means: we know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process. For example, if we took family income (x), this value would provide some useful information about how much financial support a college may offer a prospective student (y). However, the prediction would be far from perfect, since other factors play a role in financial support beyond a family's finances.

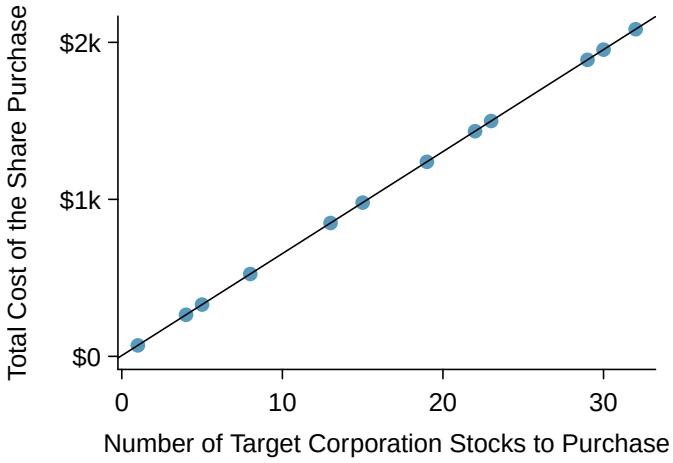


Figure 8.1: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker `TGT`, December 28th, 2018), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.

Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The values β_0 and β_1 represent the model's parameters (β is the Greek letter *beta*), and the error is represented by ε (the Greek letter *epsilon*). The parameters are estimated using data, and we write their point estimates as b_0 and b_1. When we use x to predict y, we usually call x the explanatory or **predictor** variable, and we call y the response; we also often drop the ϵ term when writing down the model since our main focus is often on the prediction of the average outcome.

It is rare for all of the data to fall perfectly on a straight line. Instead, it's more common for data to appear as a *cloud of points*, such as those examples shown in Figure 8.2. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it. In each of these examples,

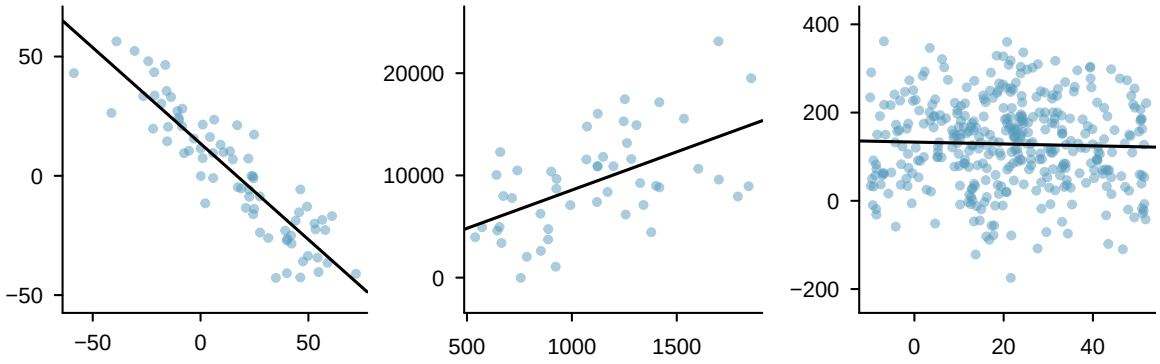


Figure 8.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

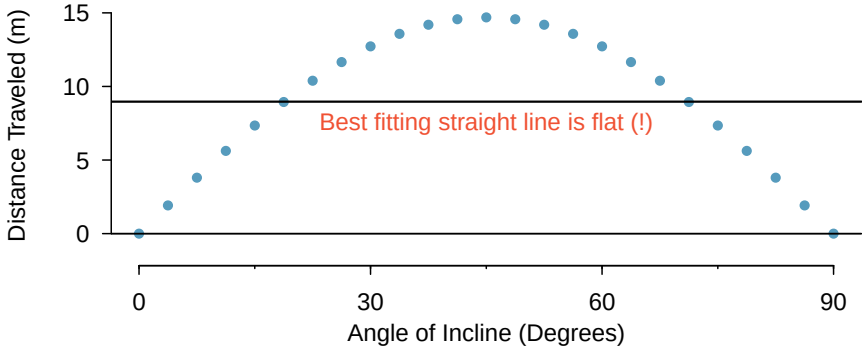


Figure 8.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

we will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less? As we move forward in this chapter, we will learn about criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

There are also cases where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 8.3 where there is a very clear relationship between the variables even though the trend is not linear. We will discuss nonlinearrends in this chapter and the next, but details of fitting nonlinear models are saved for a later course.

8.1.2 Using linear regression to predict possum head lengths

Brushtail possums are a marsupial that lives in Australia, and a photo of one is shown in Figure 8.4. Researchers captured 104 of these animals and took body measurements before releasing the animals back into the wild. We consider two of these measurements: the total length of each possum, from head to tail, and the length of each possum's head.

Figure 8.5 shows a scatterplot for the head length and total length of the possums. Each point represents a single possum from the data. The head and total length variables are associated: possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length as the predictor variable, x, to predict a possum's head length, y. We could fit the linear relationship by eye, as in Figure 8.6.

Figure 8.4: The common brushtail possum of Australia.

Photo by Greg Schechter (https://flic.kr/p/9BAFbR). CC BY 2.0 license.

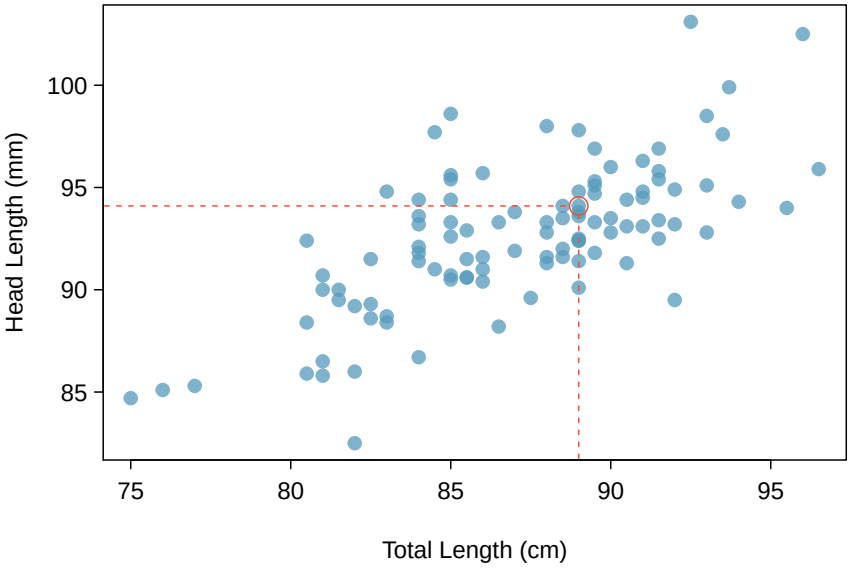


Figure 8.5: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

The equation for this line is

$$\hat{y} = 41 + 0.59x$$

A "hat" on y is used to signify that this is an estimate. We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

$$\begin{aligned}\hat{y} &= 41 + 0.59 \times 80 \\ &= 88.2\end{aligned}$$

The estimate may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

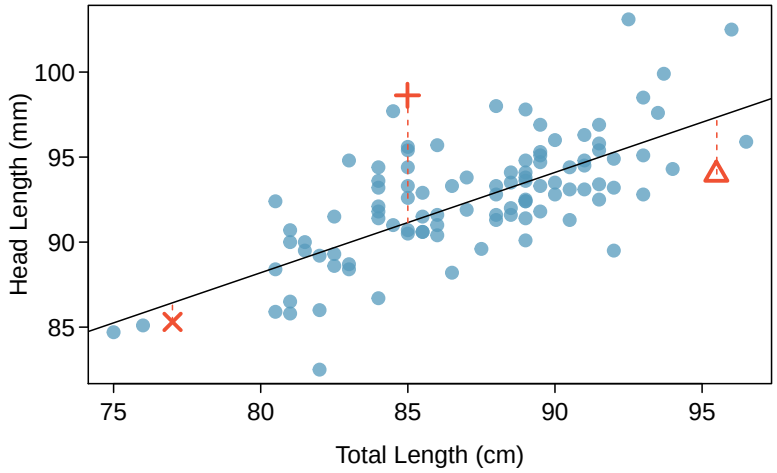


Figure 8.6: A reasonable linear model was fit to represent the relationship between head length and total length.

EXAMPLE 8.1

What other variables might help us predict the head length of a possum besides its length?

(E) Perhaps the relationship would be a little different for male possums than female possums, or perhaps it would differ for possums from one region of Australia versus another region. In Chapter 9, we'll learn about how we can include more than one predictor. Before we get there, we first need to better understand how to best build a simple linear model with one predictor.

8.1.3 Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Each observation will have a residual, and three of the residuals for the linear model we fit for the `possum` data is shown in Figure 8.6. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

Let's look closer at the three residuals featured in Figure 8.6. The observation marked by an "×" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by "△" has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for "△" is larger than that of "×" because $|-4|$ is larger than $|-1|$.

RESIDUAL: DIFFERENCE BETWEEN OBSERVED AND EXPECTED

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

EXAMPLE 8.2

The linear fit shown in Figure 8.6 is given as $\hat{y} = 41 + 0.59x$. Based on this line, formally compute the residual of the observation $(77.0, 85.3)$. This observation is denoted by "×" in Figure 8.6. Check it against the earlier visual estimate, -1.

We first compute the predicted value of point "×" based on the model:

$$\hat{y}_\times = 41 + 0.59x_\times = 41 + 0.59 \times 77.0 = 86.4$$

Next we compute the difference of the actual head length and the predicted head length:

$$e_\times = y_\times - \hat{y}_\times = 85.3 - 86.4 = -1.1$$

The model's error is $e_\times = -1.1$mm, which is very close to the visual estimate of -1mm.

GUIDED PRACTICE 8.3

If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

GUIDED PRACTICE 8.4

Compute the residuals for the "+" observation $(85.0, 98.6)$ and the "△" observation $(95.5, 94.0)$ in the figure using the linear relationship $\hat{y} = 41 + 0.59x$.[2]

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a **residual plot** such as the one shown in Figure 8.7 for the regression line in Figure 8.6. The residuals are plotted at their original horizontal locations but with the vertical coordinate as the residual. For instance, the point $(85.0, 98.6)_+$ had a residual of 7.45, so in the residual plot it is placed at $(85.0, 7.45)$. Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal.

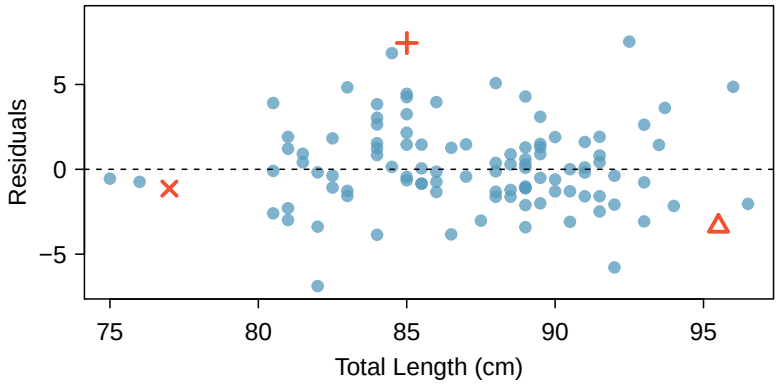


Figure 8.7: Residual plot for the model in Figure 8.6.

[1] If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative.

[2] (+) First compute the predicted value based on the model:

$$\hat{y}_+ = 41 + 0.59x_+ = 41 + 0.59 \times 85.0 = 91.15$$

Then the residual is given by

$$e_+ = y_+ - \hat{y}_+ = 98.6 - 91.15 = 7.45$$

This was close to the earlier estimate of 7.
(△) $\hat{y}_\triangle = 41 + 0.59x_\triangle = 97.3$. $e_\triangle = y_\triangle - \hat{y}_\triangle = -3.3$, close to the estimate of -4.

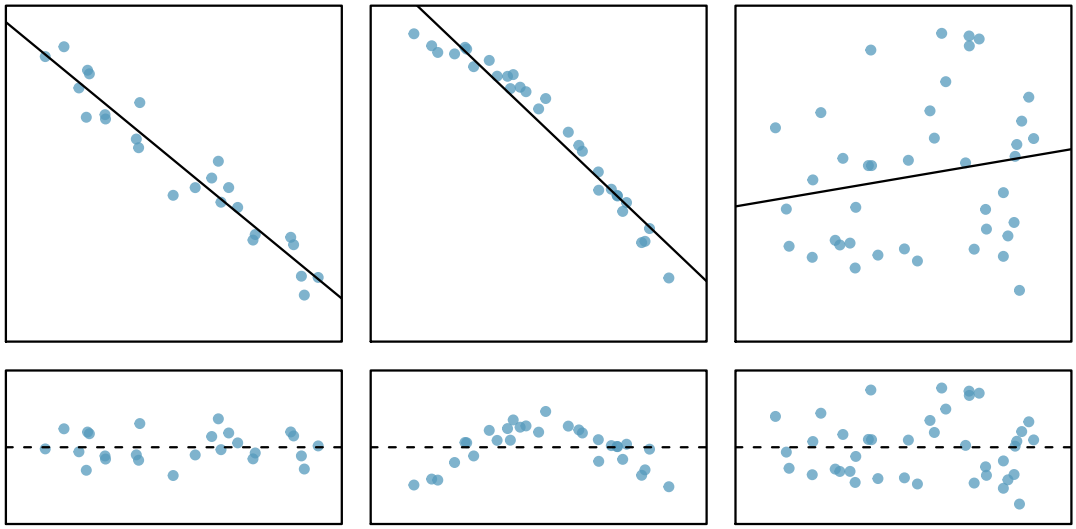


Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

EXAMPLE 8.5

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 8.8 shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

In the first data set (first column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0.

(E) The second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is statistically significant evidence that the slope parameter is different from zero. The point estimate of the slope parameter, labeled b_1, is not zero, but we might wonder if this could just be due to chance. We will address this sort of scenario in Section 8.4.

8.1.4 Describing linear relationships with correlation

We've seen plots with strong linear relationships and others with very weak linear relationships. It would be useful if we could quantify the strength of these linear relationships with a statistic.

CORRELATION: STRENGTH OF A LINEAR RELATIONSHIP

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

We can compute the correlation using a formula, just as we did with the sample mean and standard deviation. This formula is rather complex,[3] and like with other statistics, we generally perform the calculations on a computer or calculator. Figure 8.9 shows eight plots and their corresponding

[3]Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) using the formula

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$$

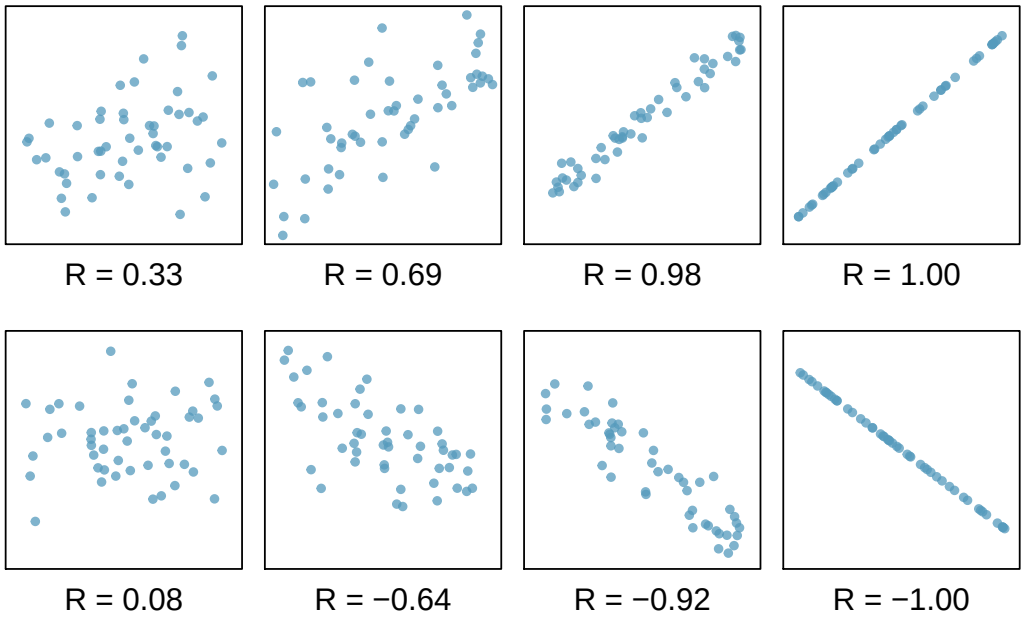


Figure 8.9: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

The correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 8.10.

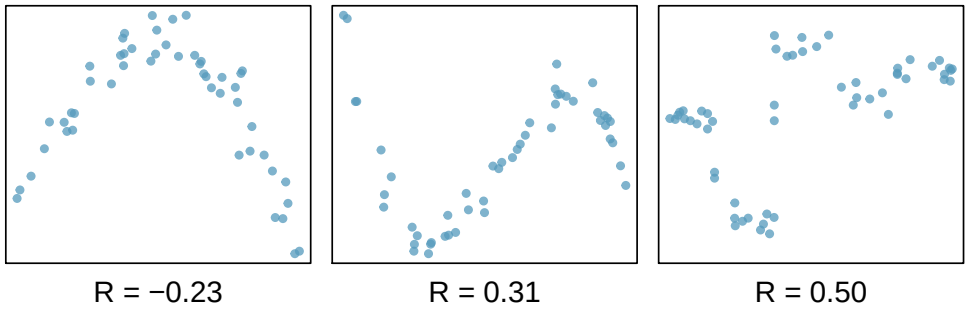


Figure 8.10: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, because the relationship is nonlinear, the correlation is relatively weak.

GUIDED PRACTICE 8.6

G No straight line is a good fit for the data sets represented in Figure 8.10. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

[4]We'll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reflect overall trends in the data.

8.2 Least squares regression

Fitting linear models by eye is open to criticism since it is based on an individual's preference. In this section, we use *least squares regression* as a more rigorous approach.

8.2.1 Gift aid for freshman at Elmhurst College

This section considers family income and gift aid data from a random sample of fifty students in the freshman class of Elmhurst College in Illinois. Gift aid is financial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 8.11 along with two linear fits. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university.

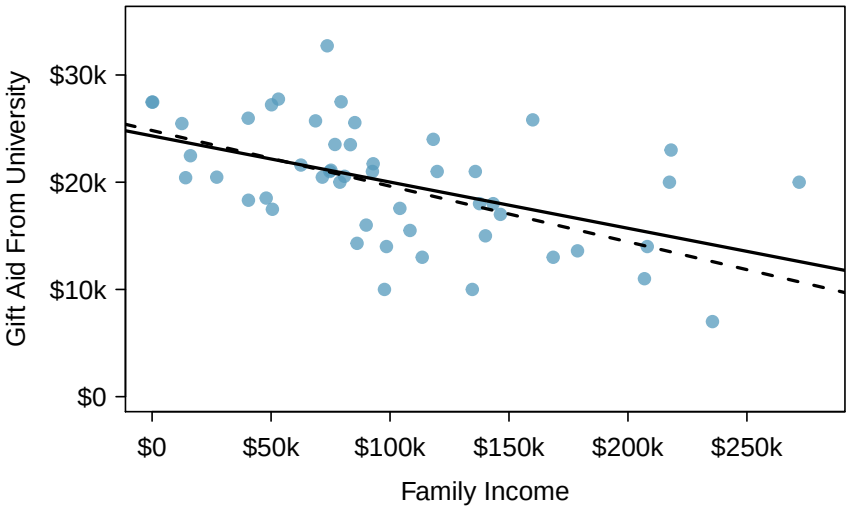


Figure 8.11: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

G **GUIDED PRACTICE 8.7**

Is the correlation positive or negative in Figure 8.11?[5]

8.2.2 An objective measure for finding the best line

We begin by thinking about what we mean by "best". Mathematically, we want a line that has small residuals. The first option that may come to mind is to minimize the sum of the residual magnitudes:

$$|e_1| + |e_2| + \cdots + |e_n|$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 8.11 demonstrates this fit can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

[5]Larger family incomes are associated with lower amounts of aid, so the correlation will be negative. Using a computer, the correlation can be computed: -0.499.

The line that minimizes this **least squares criterion** is represented as the solid line in Figure 8.11. This is commonly called the **least squares line**. The following are three possible reasons to choose this option instead of trying to minimize the sum of residual magnitudes without any squaring:

1. It is the most commonly used method.
2. Computing the least squares line is widely supported in statistical software.
3. In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.

The first two reasons are largely for tradition and convenience; the last reason explains why the least squares criterion is typically most helpful.[6]

8.2.3 Conditions for the least squares line

When fitting a least squares line, we generally require

Linearity. The data should show a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 8.12), an advanced regression method from another book or later course should be applied.

Nearly normal residuals. Generally, the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points, which we'll talk about more in Sections 8.3. An example of a residual that would be a potentially concern is shown in Figure 8.12, where one observation is clearly much further from the regression line than the others.

Constant variability. The variability of points around the least squares line remains roughly constant. An example of non-constant variability is shown in the third panel of Figure 8.12, which represents the most common pattern observed when this condition fails: the variability of y is larger when x is larger.

Independent observations. Be cautious about applying regression to **time series** data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis. An example of a data set where successive observations are not independent is shown in the fourth panel of Figure 8.12. There are also other instances where correlations within the data are important, which is further discussed in Chapter 9.

GUIDED PRACTICE 8.8

(G) Should we have concerns about applying least squares regression to the Elmhurst data in Figure 8.11?[7]

8.2.4 Finding the least squares line

For the Elmhurst data, we could write the equation of the least squares regression line as

$$\widehat{aid} = \beta_0 + \beta_1 \times family\_income$$

Here the equation is set up to predict gift aid based on a student's family income, which would be useful to students considering Elmhurst. These two values, β_0 and β_1, are the parameters of the regression line.

[6] There are applications where the sum of residual magnitudes may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

[7] The trend appears to be linear, the data fall around the line with no obvious outliers, the variance is roughly constant. These are also not time series observations. Least squares regression can be applied to these data.

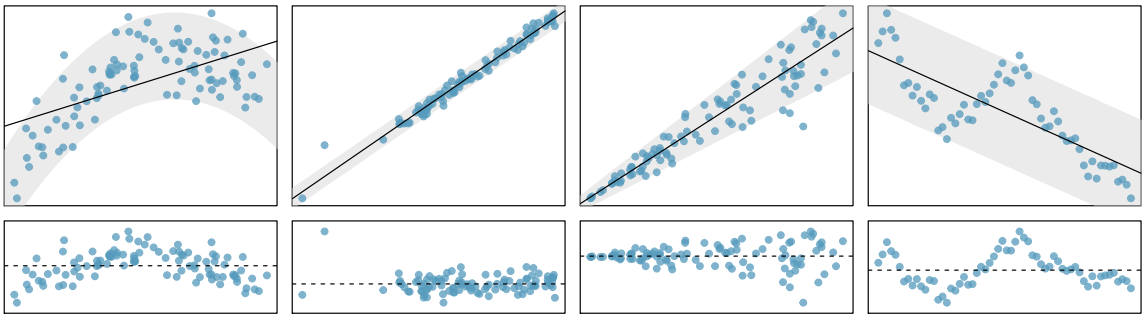


Figure 8.12: Four examples showing when the methods in this chapter are insufficient to apply to the data. First panel: linearity fails. Second panel: there are outliers, most especially one point that is very far away from the line. Third panel: the variability of the errors is related to the value of x. Fourth panel: a time series data set is shown, where successive observations are highly correlated.

As in Chapters 5, 6, and 7, the parameters are estimated using observed data. In practice, this estimation is done using a computer in the same way that other estimates, like a sample mean, can be estimated using a computer or calculator. However, we can also find the parameter estimates by applying two properties of the least squares line:

- The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} R$$

  where R is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable and response, respectively.

- If $\bar{x}$ is the sample mean of the explanatory variable and $\bar{y}$ is the sample mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line.

  Figure 8.13 shows the sample means for the family income and gift aid as \$101,780 and \$19,940, respectively. We could plot the point $(101.8, 19.94)$ on Figure 8.11 on page 231 to verify it falls on the least squares line (the solid line).

Next, we formally find the point estimates b_0 and b_1 of the parameters β_0 and β_1.

	Family Income ("x")	Gift Aid ("y")
mean	$\bar{x} =$ \$101,780	$\bar{y} =$ \$19,940
sd	$s_x =$ \$63,200	$s_y =$ \$5,460
		$R = -0.499$

Figure 8.13: Summary statistics for family income and gift aid.

GUIDED PRACTICE 8.9

G Using the summary statistics in Figure 8.13, compute the slope for the regression line of gift aid against family income.[8]

You might recall the **point-slope** form of a line from math class, which we can use to find the model fit, including the estimate of b_0. Given the slope of a line and a point on the line, (x_0, y_0), the equation for the line can be written as

$$y - y_0 = slope \times (x - x_0)$$

[8] Compute the slope using the summary statistics from Figure 8.13:

$$b_1 = \frac{s_y}{s_x} R = \frac{5{,}460}{63{,}200}(-0.499) = -0.0431$$

IDENTIFYING THE LEAST SQUARES LINE FROM SUMMARY STATISTICS

To identify the least squares line from summary statistics:

- Estimate the slope parameter, $b_1 = (s_y/s_x)R$.
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ with the point-slope equation: $y - \bar{y} = b_1(x - \bar{x})$.
- Simplify the equation, which would reveal that $b_0 = \bar{y} - b_1\bar{x}$.

EXAMPLE 8.10

Using the point (101780, 19940) from the sample means and the slope estimate $b_1 = -0.0431$ from Guided Practice 8.9, find the least-squares line for predicting aid based on family income.

Apply the point-slope equation using (101.8, 19.94) and the slope $b_1 = -0.0431$:

$$y - y_0 = b_1(x - x_0)$$
$$y - 19{,}940 = -0.0431(x - 101{,}780)$$

Expanding the right side and then adding 19,940 to each side, the equation simplifies:

$$\widehat{aid} = 24{,}320 - 0.0431 \times family\_income$$

Here we have replaced y with $\widehat{aid}$ and x with $family\_income$ to put the equation in context. The final equation should always include a "hat" on the variable being predicted, whether it is a generic "y" or a named variable like "aid".

A computer is usually used to compute the least squares line, and a summary table generated using software for the Elmhurst regression line is shown in Figure 8.14. The first column of numbers provides estimates for b_0 and b_1, respectively. These results match those from Example 8.10.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24319.3	1291.5	18.83	<0.0001
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.14: Summary of least squares fit for the Elmhurst data. Compare the parameter estimates in the first column to the results of Example 8.10.

EXAMPLE 8.11

Examine the second, third, and fourth columns in Figure 8.14. Can you guess what they represent? (If you have not reviewed any inference chapter yet, skip this Guided Practice exercise.)

We'll describe the meaning of the columns using the second row, which corresponds to β_1. The first column provides the point estimate for β_1, as we calculated in an earlier example: $b_1 = -0.0431$. The second column is a standard error for this point estimate: $SE_{b_1} = 0.0108$. The third column is a t-test statistic for the null hypothesis that $\beta_1 = 0$: $T = -3.98$. The last column is the p-value for the t-test statistic for the null hypothesis $\beta_1 = 0$ and a two-sided alternative hypothesis: 0.0002. We will get into more of these details in Section 8.4.

EXAMPLE 8.12

Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have estimated to calculate her financial aid from the university?

(E)

She may use it as an estimate, though some qualifiers on this approach are important. First, the data all come from one freshman class, and the way aid is determined by the university may change from year to year. Second, the equation will provide an imperfect estimate. While the linear equation is good at capturing the trend in the data, no individual student's aid will be perfectly predicted.

8.2.5 Interpreting regression line parameter estimates

Interpreting parameters in a regression model is often one of the most important steps in the analysis.

EXAMPLE 8.13

The intercept and slope estimates for the Elmhurst data are $b_0 = 24{,}319$ and $b_1 = -0.0431$. What do these numbers really mean?

(E)

Interpreting the slope parameter is helpful in almost any application. For each additional \$1,000 of family income, we would expect a student to receive a net difference of $\$1{,}000 \times (-0.0431) = -\$43.10$ in aid on average, i.e. \$43.10 *less*. Note that a higher family income corresponds to less aid because the coefficient of family income is negative in the model. We must be cautious in this interpretation: while there is a real association, we cannot interpret a causal connection between the variables because these data are observational. That is, increasing a student's family income may not cause the student's aid to drop. (It would be reasonable to contact the college and ask if the relationship is causal, i.e. if Elmhurst College's aid decisions are partially based on students' family income.)

The estimated intercept $b_0 = 24{,}319$ describes the average aid if a student's family had no income. The meaning of the intercept is relevant to this application since the family income for some students at Elmhurst is \$0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero.

INTERPRETING PARAMETERS ESTIMATED BY LEAST SQUARES

The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if $x = 0$ *and* the linear model is valid all the way to $x = 0$, which in many applications is not the case.

8.2.6 Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6^{th} it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

Stephen Colbert
April 6th, 2010[9]

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

[9] www.cc.com/video-clips/l4nkoq

EXAMPLE 8.14

Use the model $\widehat{aid} = 24{,}319 - 0.0431 \times family\_income$ to estimate the aid of another freshman student whose family had income of $1 million.

(E)

We want to calculate the aid for $family\_income = 1{,}000{,}000$:

$$24{,}319 - 0.0431 \times family\_income = 24{,}319 - 0.0431 \times 1{,}000{,}000 = -18{,}781$$

The model predicts this student will have -$18,781 in aid (!). However, Elmhurst College does not offer *negative aid* where they select some students to pay extra on top of tuition to attend.

Applying a model estimate to values outside of the realm of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

8.2.7 Using R^2 to describe the strength of a fit

We evaluated the strength of the linear relationship between two variables earlier using the correlation, R. However, it is more common to explain the strength of a linear fit using R^2, called **R-squared**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

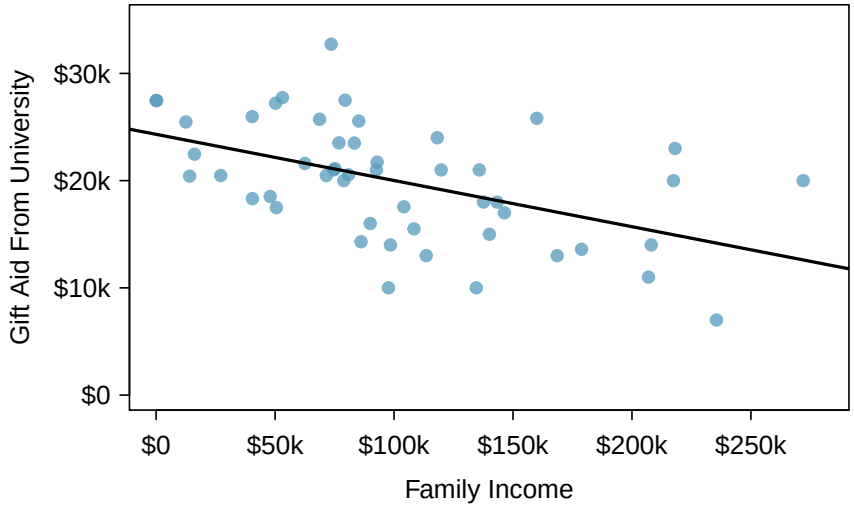


Figure 8.15: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line. For example, consider the Elmhurst data, shown in Figure 8.15. The variance of the response variable, aid received, is about $s^2_{aid} \approx 29.8$ million. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's family income. The variability in the residuals describes how much variation remains after using the model: $s^2_{RES} \approx 22.4$ million. In short, there was a reduction of

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29{,}800{,}000 - 22{,}400{,}000}{29{,}800{,}000} = \frac{7{,}500{,}000}{29{,}800{,}000} = 0.25$$

or about 25% in the data's variation by using information about family income for predicting aid using a linear model. This corresponds exactly to the R-squared value:

$$R = -0.499 \qquad\qquad R^2 = 0.25$$

GUIDED PRACTICE 8.15

If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?[10]

8.2.8 Categorical predictors with two levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*). We'll consider Ebay auctions for a video game, *Mario Kart* for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded. Here we want to predict total price based on game condition, which takes values `used` and `new`. A plot of the auction data is shown in Figure 8.16.

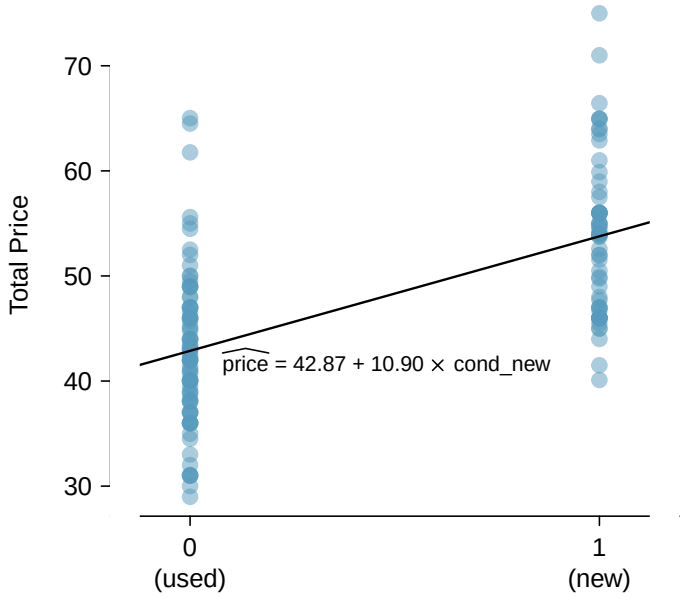


Figure 8.16: Total auction prices for the video game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `cond_new`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{price} = \beta_0 + \beta_1 \times \texttt{cond\_new}$$

The parameter estimates are given in Figure 8.17, and the model equation can be summarized as

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

For categorical predictors with just two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance. As can be seen in Figure 8.16, both of these conditions are reasonably satisfied by the auction data.

[10] About $R^2 = (-0.97)^2 = 0.94$ or 94% of the variation is explained by the linear model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	<0.0001
cond_new	10.90	1.26	8.66	<0.0001

Figure 8.17: Least squares regression summary for the final auction price against the condition of the game.

EXAMPLE 8.16

Interpret the two parameters estimated in the model for the price of *Mario Kart* in eBay auctions.

E

The intercept is the estimated price when `cond_new` takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.

INTERPRETING MODEL ESTIMATES FOR CATEGORICAL PREDICTORS

The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

We'll elaborate further on this topic in Chapter 9, where we examine the influence of many predictor variables simultaneously using multiple regression.

8.3 Types of outliers in linear regression

In this section, we identify criteria for determining which outliers are important and influential. Outliers in regression are observations that fall far from the cloud of points. These points are especially important because they can have a strong influence on the least squares line.

EXAMPLE 8.17

There are six plots shown in Figure 8.18 along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify the outliers and note how they influence the least squares line. Recall that an outlier is any point that doesn't appear to belong with the vast majority of the other points.

(1) There is one outlier far from the other points, though it only appears to slightly influence the line.

(2) There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.

(E) (3) There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well.

(4) There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

(5) There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.

(6) There is one outlier far from the cloud. However, it falls quite close to the least squares line and does not appear to be very influential.

Examine the residual plots in Figure 8.18. You will probably find that there is some trend in the main clouds of (3) and (4). In these cases, the outliers influenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

LEVERAGE

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage**.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in cases (3), (4), and (5) of Example 8.17 – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

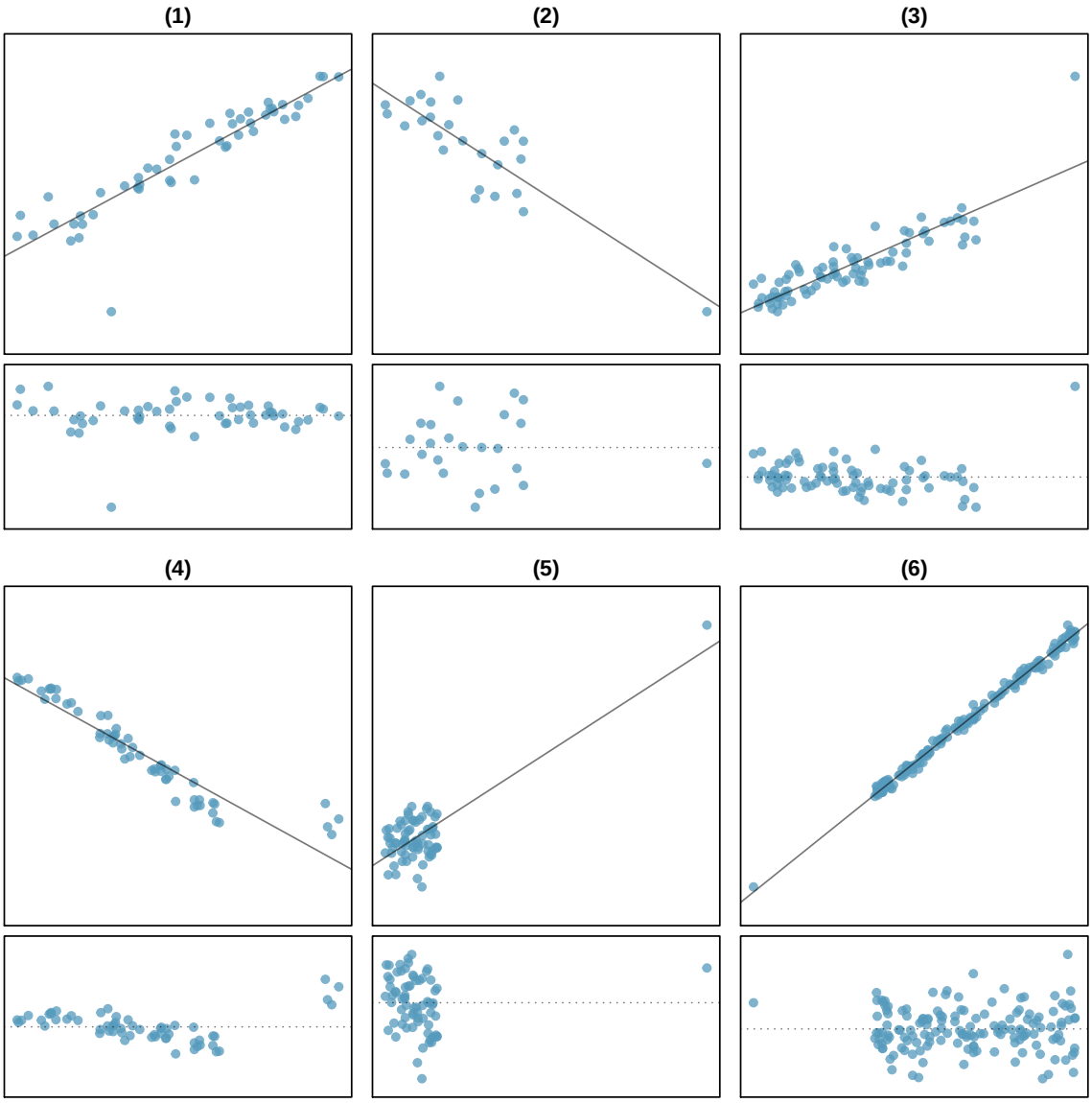


Figure 8.18: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

8.4 Inference for linear regression

In this section, we discuss uncertainty in the estimates of the slope and y-intercept for a regression line. Just as we identified standard errors for point estimates in previous chapters, we first discuss standard errors for these new estimates.

8.4.1 Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. Figure 8.19 shows these data and the least-squares regression line:

$$\begin{aligned}&\%\text{ change in House seats for President's party}\\&\quad = -7.36 - 0.89 \times (\text{unemployment rate})\end{aligned}$$

We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Republicans in 2018) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or substantial outliers. While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found.

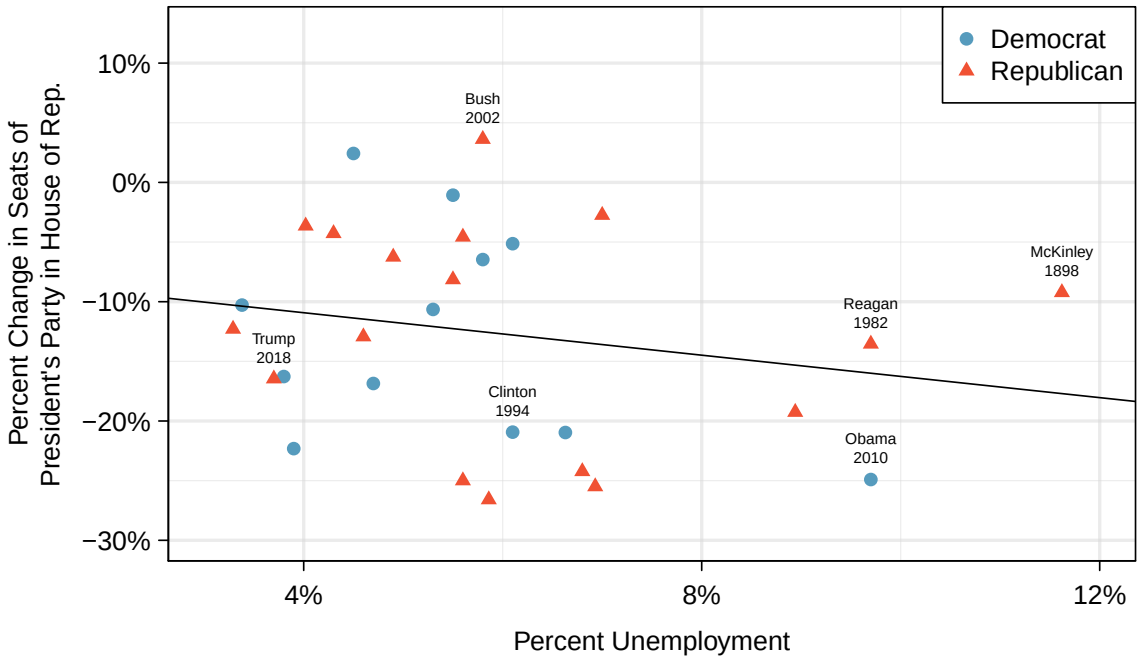


Figure 8.19: The percent change in House seats for the President's party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data.

GUIDED PRACTICE 8.18

G The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?[11]

There is a negative slope in the line shown in Figure 8.19. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the "true" linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate, where the unemployment rate is a useful predictor of the midterm election? We can frame this investigation into a statistical hypothesis test:

H_0: $\beta_1 = 0$. The true linear model has slope zero.

H_A: $\beta_1 \neq 0$. The true linear model has a slope different than zero. The unemployment is predictive of whether the President's party wins or loses seats in the House of Representatives.

We would reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter is different than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value.

8.4.2 Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b_1. We will generally label the test statistic using a T, since it follows the t-distribution.

We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Figure 8.20 shows software output for the least squares regression line in Figure 8.19. The row labeled *unemp* includes the point estimate and other hypothesis test information for the slope, which is the coefficient of the unemployment variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.3644	5.1553	-1.43	0.1646
unemp	-0.8897	0.8350	-1.07	0.2961
				$df = 27$

Figure 8.20: Output from statistical software for the regression line modeling the midterm election losses for the President's party as a response to unemployment.

EXAMPLE 8.19

What do the first and second columns of Figure 8.20 represent?

E The entries in the first column represent the least squares estimates, b_0 and b_1, and the values in the second column correspond to the standard errors of each estimate. Using the estimates, we could write the equation for the least square regression line as

$$\hat{y} = -7.3644 - 0.8897x$$

where $\hat{y}$ in this case represents the predicted change in the number of seats for the president's party, and x represents the unemployment rate.

We previously used a t-test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can

[11] We will provide two considerations. Each of these points would have very high leverage on any least-squares regression line, and years with such high unemployment may not help us understand what would happen in other years where the unemployment is only modestly high. On the other hand, these are exceptional cases, and we would be discarding important information if we exclude them from a final analysis.

compute the test statistic using the T (or Z) score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-0.8897 - 0}{0.8350} = -1.07$$

This corresponds to the third column of Figure 8.20.

EXAMPLE 8.20

Use the table in Figure 8.20 to determine the p-value for the hypothesis test.

(E)

The last column of the table gives the p-value for the two-sided hypothesis test for the coefficient of the unemployment rate: 0.2961. That is, the data do not provide convincing evidence that a higher unemployment rate has any correspondence with smaller or larger losses for the President's party in the House of Representatives in midterm elections.

INFERENCE FOR REGRESSION

We usually rely on statistical software to identify point estimates, standard errors, test statistics, and p-values in practice. However, be aware that software will not generally check whether the method is appropriate, meaning we must still verify conditions are met.

EXAMPLE 8.21

Examine Figure 8.15 on page 236, which relates the Elmhurst College aid and student family income. How sure are you that the slope is statistically significantly different from zero? That is, do you think a formal hypothesis test would reject the claim that the true slope of the line should be zero?

(E)

While the relationship between the variables is not perfect, there is an evident decreasing trend in the data. This suggests the hypothesis test will reject the null claim that the slope is zero.

GUIDED PRACTICE 8.22

Figure 8.21 shows statistical software output from fitting the least squares regression line shown in Figure 8.15. Use this output to formally evaluate the following hypotheses.[12]

(G)

H_0: The true coefficient for family income is zero.

H_A: The true coefficient for family income is not zero.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24319.3	1291.5	18.83	<0.0001
family_income	-0.0431	0.0108	-3.98	0.0002
				$df = 48$

Figure 8.21: Summary of least squares fit for the Elmhurst College data.

[12] We look in the second row corresponding to the family income variable. We see the point estimate of the slope of the line is -0.0431, the standard error of this estimate is 0.0108, and the t-test statistic is $T = -3.98$. The p-value corresponds exactly to the two-sided test we are interested in: 0.0002. The p-value is so small that we reject the null hypothesis and conclude that family income and financial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated and the true slope parameter is indeed less than 0, just as we believed in Example 8.21.

8.4.3 Confidence interval for a coefficient

Similar to how we can conduct a hypothesis test for a model coefficient using regression output, we can also construct a confidence interval for that coefficient.

EXAMPLE 8.23

Compute the 95% confidence interval for the `family_income` coefficient using the regression output from Table 8.21.

(E)

The point estimate is -0.0431 and the standard error is $SE = 0.0108$. When constructing a confidence interval for a model coefficient, we generally use a t-distribution. The degrees of freedom for the distribution are noted in the regression output, $df = 48$, allowing us to identify $t^\star_{48} = 2.01$ for use in the confidence interval.

We can now construct the confidence interval in the usual way:

$$\text{point estimate} \pm t^\star_{48} \times SE \quad \rightarrow \quad -0.0431 \pm 2.01 \times 0.0108 \quad \rightarrow \quad (-0.0648, -0.0214)$$

We are 95% confident that the true coefficient for the `family_income` variable is between -0.0648 and -0.0214.

CONFIDENCE INTERVALS FOR COEFFICIENTS

Confidence intervals for model coefficients can be computed using the t-distribution:

$$b_i \ \pm \ t^\star_{df} \times SE_{b_i}$$

where $t^\star_{df}$ is the appropriate t-value corresponding to the confidence level and model degrees of freedom, $df = n - k - 1$.

Chapter 9

Multiple and logistic regression

The principles of simple linear regression lay the foundation for more sophisticated regression methods used in a wide range of challenging settings. In Chapter 9, we explore multiple regression, which introduces the possibility of more than one predictor, and logistic regression, a technique for predicting categorical outcomes with two possible categories.

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

9.1 Introduction to multiple regression

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider data about loans from the peer-to-peer lender, Lending Club, which is a data set we first encountered in Chapters 1 and 2. The loan data includes terms of the loan as well as information about the borrower. The outcome variable we would like to better understand is the interest rate assigned to the loan. For instance, all other characteristics held constant, does it matter how much debt someone already has? Does it matter if their income has been verified? Multiple regression will help us answer these and other questions.

The data set `loans` includes results from 10,000 loans, and we'll be looking at a subset of the available variables, some of which will be new from those we saw in earlier chapters. The first six observations in the data set are shown in Figure 9.1, and descriptions for each variable are shown in Figure 9.2. Notice that the past bankruptcy variable (`bankruptcy`) is an indicator variable, where it takes the value 1 if the borrower had a past bankruptcy in their record and 0 if not. Using an indicator variable in place of a category name allows for these variables to be directly used in regression. Two of the other variables are categorical (`income_ver` and `issued`), each of which can take one of a few different non-numerical values; we'll discuss how these are handled in the model in Section 9.1.1.

	interest_rate	income_ver	debt_to_income	credit_util	bankruptcy	term	issued	credit_checks
1	14.07	verified	18.01	0.55	0	60	Mar2018	6
2	12.61	not	5.04	0.15	1	36	Feb2018	1
3	17.09	source_only	21.15	0.66	0	36	Feb2018	4
4	6.72	not	10.16	0.20	0	36	Jan2018	0
5	14.07	verified	57.96	0.75	0	36	Mar2018	7
6	6.72	not	6.46	0.09	0	36	Jan2018	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 9.1: First six rows from the `loans` data set.

variable	description
`interest_rate`	Interest rate for the loan.
`income_ver`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `verified`, `source_only`, and `not`.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in her record. This variable takes a value of `1` if the answer is "yes" and `0` if the answer is "no".
`term`	The length of the loan, in months.
`issued`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Figure 9.2: Variables and their descriptions for the `loans` data set.

9.1.1 Indicator and categorical variables as predictors

Let's start by fitting a linear regression model for interest rate with a single predictor indicating whether or not a person has a bankruptcy in their record:

$$\widehat{rate} = 12.33 + 0.74 \times bankruptcy$$

Results of this model are shown in Figure 9.3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3380	0.0533	231.49	<0.0001
bankruptcy	0.7368	0.1529	4.82	<0.0001
				$df = 9998$

Figure 9.3: Summary of a linear model for predicting interest rate based on whether the borrower has a bankruptcy in their record.

EXAMPLE 9.1

Interpret the coefficient for the past bankruptcy variable in the model. Is this coefficient significantly different from 0?

E The `bankruptcy` variable takes one of two values: 1 when the borrower has a bankruptcy in their history and 0 otherwise. A slope of 0.74 means that the model predicts a 0.74% higher interest rate for those borrowers with a bankruptcy in their record. (See Section 8.2.8 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Figure 9.3, we can see that the p-value for `bankruptcy` is very close to zero, indicating there is strong evidence the coefficient is different from zero when using this simple one-predictor model.

Suppose we had fit a model using a 3-level categorical variable, such as `income_ver`. The output from software is shown in Figure 9.4. This regression output provides multiple rows for the `income_ver` variable. Each row represents the relative difference for each level of `income_ver`. However, we are missing one of the levels: `not` (for *not verified*). The missing level is called the **reference level**, and it represents the default level that other levels are measured against.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.0995	0.0809	137.18	<0.0001
income_ver: *source_only*	1.4160	0.1107	12.79	<0.0001
income_ver: *verified*	3.2543	0.1297	25.09	<0.0001
				$df = 9998$

Figure 9.4: Summary of a linear model for predicting interest rate based on whether the borrower's income source and amount has been verified. This predictor has three levels, which results in 2 rows in the regression output.

EXAMPLE 9.2

How would we write an equation for this regression model?

E The equation for the regression model may be written as a model with two predictors:

$$\widehat{rate} = 11.10 + 1.42 \times \texttt{income\_ver}_{\texttt{source\_only}} + 3.25 \times \texttt{income\_ver}_{\texttt{verified}}$$

We use the notation $\texttt{variable}_{\texttt{level}}$ to represent **indicator variables** for when the categorical variable takes a particular value. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ would take a value of 1 if `income_ver` was `source_only` for a loan, and it would take a value of 0 otherwise. Likewise, $\texttt{income\_ver}_{\texttt{verified}}$ would take a value of 1 if `income_ver` took a value of `verified` and 0 if it took any other value.

The notation used in Example 9.2 may feel a bit confusing. Let's figure out how to use the equation for each level of the `income_ver` variable.

EXAMPLE 9.3

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source and amount are both unverified.

When `income_ver` takes a value of `not`, then both indicator functions in the equation from Example 9.2 are set to zero:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 0 \\ &= 11.10\end{aligned}$$

The average interest rate for these borrowers is 11.1%. Because the `not` level does not have its own coefficient and it is the reference value, the indicators for the other levels for this variable all drop out.

EXAMPLE 9.4

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source is verified but the amount is not.

When `income_ver` takes a value of `source_only`, then the corresponding indicator function is 1 while the other ($1_{verified}$) is 0:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 1 + 3.25 \times 0 \\ &= 12.52\end{aligned}$$

The average interest rate for these borrowers is 12.52%.

GUIDED PRACTICE 9.5

G

Compute the average interest rate for borrowers whose income source and amount are both verified.[1]

PREDICTORS WITH SEVERAL CATEGORIES

When fitting a regression model with a categorical variable that has k levels where $k > 2$, software will provide a coefficient for $k - 1$ of those levels. For the last level that does not receive a coefficient, this is the **reference level**, and the coefficients listed for the other levels are all considered relative to this reference level.

GUIDED PRACTICE 9.6

G

Interpret the coefficients in the `income_ver` model.[2]

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we'd think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding

[1] When `income_ver` takes a value of `verified`, then the corresponding indicator function is 1 while the other (1_{source_only}) is 0:

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 1 \\ &= 14.35\end{aligned}$$

The average interest rate for these borrowers is 14.35%.

[2] Each of the coefficients gives the incremental interest rate for the corresponding level relative to the `not` level, which is the reference level. For example, for a borrower whose income source and amount have been verified, the model predicts that they will have a 3.25% higher interest rate than a borrower who has not had their income source or amount verified.

variables that we didn't account for. For example, perhaps lender require borrowers with poor credit to verify their income. That is, verifying income in our data set might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate. (What other confounding variables might explain this counter-intuitive relationship suggested by the model?)

GUIDED PRACTICE 9.7

G How much larger of an interest rate would we expect for a borrower who has verified their income source and amount vs a borrower whose income source has only been verified?[3]

9.1.2 Including and assessing many variables in a model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling. For example, we might like to use the full context of borrower to predict the interest rate they receive rather than using a single variable. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression on observational data, such models are a common first step in gaining insights or providing some evidence of a causal connection.

We want to construct a model that accounts for not only for any past bankruptcy or whether the borrower had their income source or amount verified, but simultaneously accounts for all the variables in the data set: `income_ver`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `issued`, and `credit_checks`.

$$\begin{aligned}\widehat{\texttt{rate}} = \beta_0 &+ \beta_1 \times \texttt{income\_ver}_{\texttt{source\_only}} + \beta_2 \times \texttt{income\_ver}_{\texttt{verified}} + \beta_3 \times \texttt{debt\_to\_income} \\ &+ \beta_4 \times \texttt{credit\_util} + \beta_5 \times \texttt{bankruptcy} + \beta_6 \times \texttt{term} \\ &+ \beta_7 \times \texttt{issued}_{\texttt{Jan2018}} + \beta_8 \times \texttt{issued}_{\texttt{Mar2018}} + \beta_9 \times \texttt{credit\_checks}\end{aligned}$$

This equation represents a holistic approach for modeling all of the variables simultaneously. Notice that there are two coefficients for `income_ver` and also two coefficients for `issued`, since both are 3-level categorical variables.

We estimate the parameters β_0, β_1, β_2, ..., β_9 in the same way as we did in the case of a single predictor. We select b_0, b_1, b_2, ..., b_9 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{10000}^2 = \sum_{i=1}^{10000} e_i^2 = \sum_{i=1}^{10000} (y_i - \hat{y}_i)^2 \tag{9.8}$$

where y_i and $\hat{y}_i$ represent the observed interest rates and their estimated values according to the model, respectively. 10,000 residuals are calculated, one for each observation. We typically use a computer to minimize the sum of squares and compute point estimates, as shown in the sample output in Figure 9.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

MULTIPLE REGRESSION MODEL

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors. We always estimate the β_i parameters using statistical software.

[3] Relative to the `not` category, the `verified` category has an interest rate of 3.25% higher, while the `source_only` category is only 1.42% higher. Thus, `verified` borrowers will tend to get an interest rate about $3.25\% - 1.42\% = 1.83\%$ higher than `source_only` borrowers.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001
				$df = 9990$

Figure 9.5: Output for the regression model, where `interest_rate` is the outcome and the variables listed are the predictors.

EXAMPLE 9.9

Write out the regression model using the point estimates from Figure 9.5. How many predictors are there in this model?

The fitted model for the interest rate is given by:

(E)

$$\widehat{\texttt{rate}} = 1.925 + 0.975 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.537 \times \texttt{income\_ver}_{\texttt{verified}} + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.386 \times \texttt{bankruptcy} + 0.154 \times \texttt{term} + 0.028 \times \texttt{issued}_{\texttt{Jan2018}} - 0.040 \times \texttt{issued}_{\texttt{Mar2018}} + 0.228 \times \texttt{credit\_checks}$$

If we count up the number of predictor coefficients, we get the *effective* number of predictors in the model: $k = 9$. Notice that the `issued` categorical predictor counts as two, once for the two levels shown in the model. In general, a categorical predictor with p different levels will be represented by $p - 1$ terms in a multiple regression model.

(G)

GUIDED PRACTICE 9.10

What does β_4, the coefficient of variable `credit_util`, represent? What is the point estimate of β_4?[4]

EXAMPLE 9.11

Compute the residual of the first observation in Figure 9.1 on page 247 using the equation identified in Guided Practice 9.9.

(E)

To compute the residual, we first need the predicted value, which we compute by plugging values into the equation from Example 9.9. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ takes a value of 0, $\texttt{income\_ver}_{\texttt{verified}}$ takes a value of 1 (since the borrower's income source and amount were verified), `debt_to_income` was 18.01, and so on. This leads to a prediction of $\widehat{rate}_1 = 18.09$. The observed interest rate was 14.07%, which leads to a residual of $e_1 = 14.07 - 18.09 = -4.02$.

[4] β_4 represents the change in interest rate we would expect if someone's credit utilization was 0 and went to 1, all other factors held even. The point estimate is $b_4 = 4.90\%$.

EXAMPLE 9.12

We estimated a coefficient for `bankruptcy` in Section 9.1.1 of $b_4 = 0.74$ with a standard error of $SE_{b_1} = 0.15$ when using simple linear regression. Why is there be a difference between that estimate and the estimated coefficient of 0.39 in the multiple regression setting?

(E) If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `interest_rate` and predictor `bankruptcy` using simple linear regression, we were unable to control for other variables like whether the borrower had her income verified, the borrower's debt-to-income ratio, and other variables. That original model was constructed in a vacuum and did not consider the full context. When we all of the variables, underlying and unintentional bias that was missed by these other variables is reduced or eliminated. Of course, bias can still exist from other confounding variables.

Example 9.12 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** (pronounced as *co-linear*) when they are correlated, and this collinearity complicates model estimation. While it is impossible to prevent collinearity from arising in observational data, experiments are usually designed to prevent predictors from being collinear.

GUIDED PRACTICE 9.13

(G) The estimated value of the intercept is 1.925, and one might be tempted to make some interpretation of this coefficient, such as, it is the model's predicted price when each of the variables take value zero: income source is not verified, the borrower has no debt (debt-to-income and credit utilization are zero), and so on. Is this reasonable? Is there any value gained by making this interpretation?[5]

9.1.3 Adjusted R^2 as a better tool for multiple regression

We first used R^2 in Section 8.2 to determine the amount of variability in the response that was explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

where e_i represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can make it even more informative when comparing models.

GUIDED PRACTICE 9.14

(G) The variance of the residuals for the model given in Guided Practice 9.9 is 18.53, and the variance of the total price in all the auctions is 25.01. Calculate R^2 for this model.[6]

This strategy for estimating R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is a less estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R^2.

[5] Many of the variables do take a value 0 for at least one data point, and for those variables, it is reasonable. However, one variable never takes a value of zero: `term`, which describes the length of the loan, in months. If `term` is set to zero, then the loan must be paid back immediately; the borrower must give the money back as soon as she receives it, which means it is not a real loan. Ultimately, the interpretation of the intercept in this setting is not insightful.

[6] $R^2 = 1 - \frac{18.53}{25.01} = 0.2591$.

ADJUSTED R^2 AS A TOOL FOR MODEL ASSESSMENT

The **adjusted R^2** is computed as

$$R^2_{adj} = 1 - \frac{s^2_{\text{residuals}}/(n-k-1)}{s^2_{\text{outcome}}/(n-1)} = 1 - \frac{s^2_{\text{residuals}}}{s^2_{\text{outcome}}} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model. Remember that a categorical predictor with p levels will contribute $p-1$ to the number of variables in the model.

Because k is never negative, the adjusted R^2 will be smaller – often times just a little smaller – than the unadjusted R^2. The reasoning behind the adjusted R^2 lies in the **degrees of freedom** associated with each variance, which is equal to $n-k-1$ for the multiple regression context. If we were to make predictions for *new data* using our current model, we would find that the unadjusted R^2 would tend to be slightly overly optimistic, while the adjusted R^2 formula helps correct this bias.

GUIDED PRACTICE 9.15

G There were $n = 10000$ auctions in the `loans` data set and $k = 9$ predictor variables in the model. Use n, k, and the variances from Guided Practice 9.14 to calculate R^2_{adj} for the interest rate model.[7]

GUIDED PRACTICE 9.16

G Suppose you added another predictor to the model, but the variance of the errors $Var(e_i)$ didn't go down. What would happen to the R^2? What would happen to the adjusted R^2?[8]

Adjusted R^2 could have been used in Chapter 8. However, when there is only $k = 1$ predictors, adjusted R^2 is very close to regular R^2, so this nuance isn't typically important when the model has only one predictor.

[7] $R^2_{adj} = 1 - \frac{18.53}{25.01} \times \frac{10000-1}{1000-9-1} = 0.2584$. While the difference is very small, it will be important when we fine tune the model in the next section.

[8] The unadjusted R^2 would stay the same and the adjusted R^2 would go down.

9.2 Model selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate variables from the model that are found to be less important.

In practice, the model that includes all available explanatory variables is often referred to as the **full model**. The full model may not be the best model, and if it isn't, we want to identify a smaller model that is preferable.

9.2.1 Identifying variables in the model that may not be helpful

Adjusted R^2 describes the strength of a model fit, and it is a useful tool for evaluating which predictors are adding value to the model, where *adding value* means they are (likely) improving the accuracy in predicting future outcomes.

Let's consider two models, which are shown in Tables 9.6 and 9.7. The first table summarizes the full model since it includes all predictors, while the second does not include the `issued` variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001
$R^2_{adj} = 0.25843$				$df = 9990$

Figure 9.6: The fit for the full regression model, including the adjusted R^2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9213	0.1982	9.69	<0.0001
income_ver: *source_only*	0.9740	0.0991	9.83	<0.0001
income_ver: *verified*	2.5355	0.1172	21.64	<0.0001
debt_to_income	0.0211	0.0029	7.19	<0.0001
credit_util	4.8958	0.1619	30.25	<0.0001
bankruptcy	0.3869	0.1324	2.92	0.0035
term	0.1537	0.0039	38.97	<0.0001
credit_checks	0.2283	0.0182	12.51	<0.0001
$R^2_{adj} = 0.25854$				$df = 9992$

Figure 9.7: The fit for the regression model after dropping the `issued` variable.

EXAMPLE 9.17

Which of the two models is better?

E

We compare the adjusted R^2 of each model to determine which to choose. Since the first model has an R^2_{adj} smaller than the R^2_{adj} of the second model, we prefer the second model to the first.

Will the model without `issued` be better than the model with `issued`? We cannot know for sure, but based on the adjusted R^2, this is our best assessment.

9.2.2 Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward elimination* and *forward selection*. These techniques are often referred to as **stepwise** model selection strategies, because they add or delete one variable at a time as they "step" through the candidate predictors.

Backward elimination starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until we cannot improve the adjusted R^2. The strategy within each elimination step is to eliminate the variable that leads to the largest improvement in adjusted R^2.

EXAMPLE 9.18

Results corresponding to the *full model* for the `loans` data are shown in Figure 9.6. How should we proceed under the backward elimination strategy?

Our baseline adjusted R^2 from the full model is $R^2_{adj} = 0.25843$, and we need to determine whether dropping a predictor will improve the adjusted R^2. To check, we fit models that each drop a different predictor, and we record the adjusted R^2:

Exclude ...				
	`income_ver` $R^2_{adj} = 0.22380$	`debt_to_income` $R^2_{adj} = 0.25468$	`credit_util` $R^2_{adj} = 0.19063$	`bankruptcy` $R^2_{adj} = 0.25787$
	`term` $R^2_{adj} = 0.14581$	`issued` $R^2_{adj} = 0.25854$	`credit_checks` $R^2_{adj} = 0.24689$	

The model without `issued` has the highest adjusted R^2 of 0.25854, higher than the adjusted R^2 for the full model. Because eliminating `issued` leads to a model with a higher adjusted R^2, we drop `issued` from the model.

(E) Since we eliminated a predictor from the model in the first step, we see whether we should eliminate any additional predictors. Our baseline adjusted R^2 is now $R^2_{adj} = 0.25854$. We now fit new models, which consider eliminating each of the remaining predictors in addition to `issued`:

Exclude `issued` and ...			
	`income_ver` $R^2_{adj} = 0.22395$	`debt_to_income` $R^2_{adj} = 0.25479$	`credit_util` $R^2_{adj} = 0.19074$
	`bankruptcy` $R^2_{adj} = 0.25798$	`term` $R^2_{adj} = 0.14592$	`credit_checks` $R^2_{adj} = 0.24701$

None of these models lead to an improvement in adjusted R^2, so we do not eliminate any of the remaining predictors. That is, after backward elimination, we are left with the model that keeps all predictors except `issued`, which we can summarize using the coefficients from Figure 9.7:

$$\begin{aligned}\widehat{rate} = \; & 1.921 + 0.974 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.535 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.387 \times \texttt{bankruptcy} \\ & + 0.154 \times \texttt{term} + 0.228 \times \texttt{credit\_check}\end{aligned}$$

The **forward selection** strategy is the reverse of the backward elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that improve the model (as measured by adjusted R^2).

EXAMPLE 9.19

Construct a model for the `loans` data set using the forward selection strategy.

We start with the model that includes no variables. Then we fit each of the possible models with just one variable. That is, we fit the model including just `income_ver`, then the model including just `debt_to_income`, then a model with just `credit_util`, and so on. Then we examine the adjusted R^2 for each of these models:

Add ...	`income_ver`	`debt_to_income`	`credit_util`	`bankruptcy`
	$R^2_{adj} = 0.05926$	$R^2_{adj} = 0.01946$	$R^2_{adj} = 0.06452$	$R^2_{adj} = 0.00222$
	`term`	`issued`	`credit_checks`	
	$R^2_{adj} = 0.12855$	$R^2_{adj} = -¡0.00018$	$R^2_{adj} = 0.01711$	

In this first step, we compare the adjusted R^2 against a baseline model that has no predictors. The no-predictors model always has $R^2_{adj} = 0$. The model with one predictor that has the largest adjusted R^2 is the model with the `term` predictor, and because this adjusted R^2 is larger than the adjusted R^2 from the model with no predictors ($R^2_{adj} = 0$), we will add this variable to our model.

We repeat the process again, this time considering 2-predictor models where one of the predictors is `term` and with a new baseline of $R^2_{adj} = 0.12855$:

E

Add `term` and ...	`income_ver`	`debt_to_income`	`credit_util`
	$R^2_{adj} = 0.16851$	$R^2_{adj} = 0.14368$	$R^2_{adj} = 0.20046$
	`bankruptcy`	`issued`	`credit_checks`
	$R^2_{adj} = 0.13070$	$R^2_{adj} = 0.12840$	$R^2_{adj} = 0.14294$

The best second predictor, `credit_util`, has a higher adjusted R^2 (0.20046) than the baseline (0.12855), so we also add `credit_util` to the model.

Since we have again added a variable to the model, we continue and see whether it would be beneficial to add a third variable:

Add `term`, `credit_util`, and ...	`income_ver`	`debt_to_income`	
	$R^2_{adj} = 0.24183$	$R^2_{adj} = 0.20810$	
	`bankruptcy`	`issued`	`credit_checks`
	$R^2_{adj} = 0.20169$	$R^2_{adj} = 0.20031$	$R^2_{adj} = 0.21629$

The model adding `income_ver` improved adjusted R^2 (0.24183 to 0.20046), so we add `income_ver` to the model.

We continue on in this way, next adding `debt_to_income`, then `credit_checks`, and `bankruptcy`. At this point, we come again to the `issued` variable: adding this variable leads to $R^2_{adj} = 0.25843$, while keeping all the other variables but excluding `issued` leads to a higher $R^2_{adj} = 0.25854$. This means we do not add `issued`. In this example, we have arrived at the same model that we identified from backward elimination.

MODEL SELECTION STRATEGIES

Backward elimination begins with the model having the largest number of predictors and eliminates variables one-by-one until we are satisfied that all remaining variables are important to the model. Forward selection starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

Backward elimination and forward selection sometimes arrive at different final models. If trying both techniques and this happens, it's common to choose the model with the larger R^2_{adj}.

9.2.3 The p-value approach, an alternative to adjusted R^2

The p-value may be used as an alternative to R^2_{adj} for model selection:

Backward elimination with the p-value approach. In backward elimination, we would identify the predictor corresponding to the largest p-value. If the p-value is above the significance level, usually $\alpha = 0.05$, then we would drop that variable, refit the model, and repeat the process. If the largest p-value is less than $\alpha = 0.05$, then we would not eliminate any predictors and the current model would be our best-fitting model.

Forward selection with the p-value approach. In forward selection with p-values, we reverse the process. We begin with a model that has no predictors, then we fit a model for each possible predictor, identifying the model where the corresponding predictor's p-value is smallest. If that p-value is smaller than $\alpha = 0.05$, we add it to the model and repeat the process, considering whether to add more variables one-at-a-time. When none of the remaining predictors can be added to the model and have a p-value less than 0.05, then we stop adding variables and the current model would be our best-fitting model.

GUIDED PRACTICE 9.20

Examine Figure 9.7 on page 254, which considers the model including all variables except the variable for the month the loan was issued. If we were using the p-value approach with backward elimination and we were considering this model, which of these variables would be up for elimination? Would we drop that variable, or would we keep it in the model?[9]

While the adjusted R^2 and p-value approaches are similar, they sometimes lead to different models, with the R^2_{adj} approach tending to include more predictors in the final model.

ADJUSTED R^2 VS P-VALUE APPROACH

When the sole goal is to improve prediction accuracy, use R^2_{adj}. This is commonly the case in machine learning applications.

When we care about understanding which variables are statistically significant predictors of the response, or if there is interest in producing a simpler model at the potential cost of a little prediction accuracy, then the p-value approach is preferred.

Regardless of whether you use R^2_{adj} or the p-value approach, or if you use the backward elimination of forward selection strategy, our job is not done after variable selection. We must still verify the model conditions are reasonable.

[9] The `bankruptcy` predictor is up for elimination since it has the largest p-value. However, since that p-value is smaller than 0.05, we would still keep it in the model.

9.3 Checking model assumptions using graphs

Multiple regression methods using the model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

generally depend on the following four assumptions:

1. the residuals of the model are nearly normal (less important for larger data sets),
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

9.3.1 Diagnostic plots

Diagnostic plots can be used to check each of these assumptions. We will consider the model from the Lending Club loans data, and check whether there are any notable concerns: [Still must update the next equation.]

$$\begin{aligned}\widehat{rate} = \; & 1.921 + 0.974 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.535 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.387 \times \texttt{bankruptcy} \\ & + 0.154 \times \texttt{term} + 0.228 \times \texttt{credit\_check}\end{aligned}$$

Check for outliers. In theory, the distribution of the residuals should be nearly normal; in practice, normality can be relaxed for most applications. Instead, we examine a histogram of the residuals to check if there are any outliers: Figure 9.8 is a histogram of these outliers. Since this is a very large data set, only particularly extreme observations would be a concern in this particular case. There are no extreme observations that might cause a concern.

If we intended to construct what are called **prediction intervals** for future observations, we would be more strict and require the residuals to be nearly normal. Prediction intervals are further discussed in an online extra on the OpenIntro website:

www.openintro.org/d?id=stat_extra_linear_regression_supp

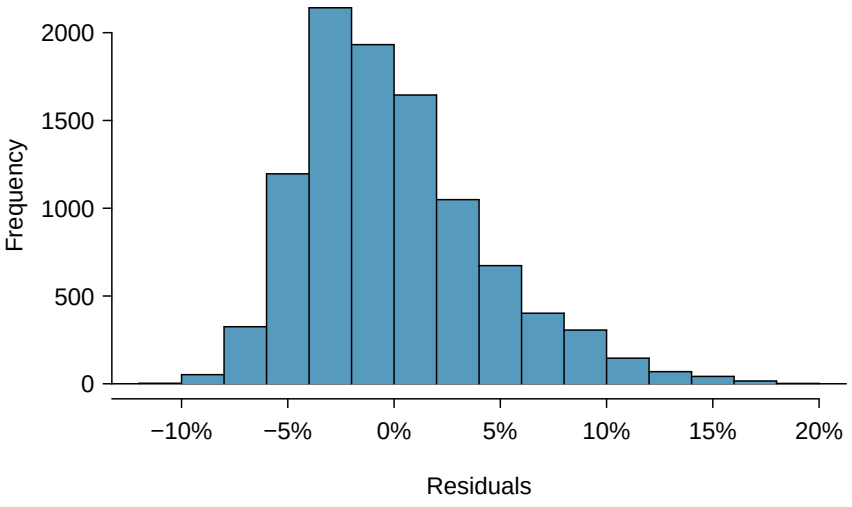


Figure 9.8: A histogram of the residuals.

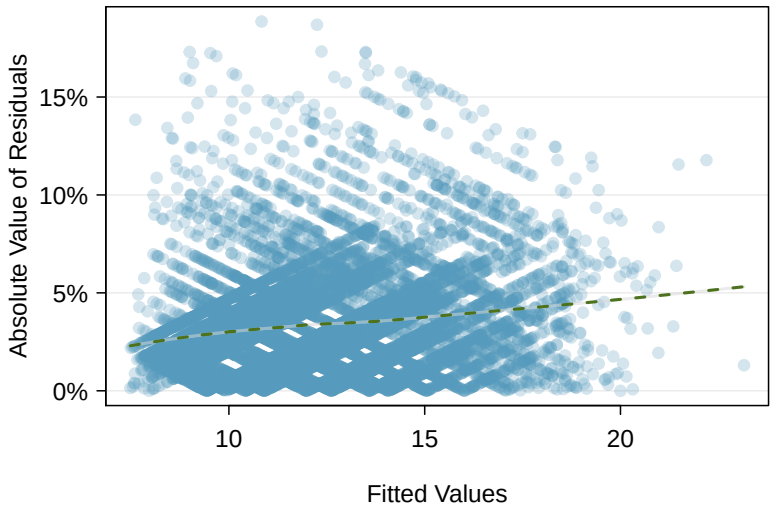


Figure 9.9: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

Absolute values of residuals against fitted values. A plot of the absolute value of the residuals against their corresponding fitted values ($\hat{y}_i$) is shown in Figure 9.9. This plot is helpful to check the condition that the variance of the residuals is approximately constant, and a smoothed line has been added to represent the approximate trend in this plot. There is more evident variability for fitted values that are larger, which we'll discuss further.

Residuals in order of their data collection. This type of plot can be helpful when observations were collected in a sequence. Such a plot is helpful in identifying any connection between cases that are close to one another. The loans in this data set were issued over a 3 month period, and the month the loan was issued was not found to be important, suggesting this is not a concern for this data set. In cases where a data set does show some pattern for this check, **time series** methods may be useful.

Residuals against each predictor variable. We consider a plot of the residuals against each of the predictors in Figure 9.10. For those instances where there are only 2-3 groups, box plots are shown. For the numerical outcomes, a smoothed line has been fit to the data to make it easier to review. Ultimately, we are looking for any notable change in variability between groups or pattern in the data.

Here are the things of importance from these plots:

- There is some minor differences in variability between the verified income groups.
- There is a very clear pattern for the debt-to-income variable. What also stands out is that this variable is very strongly right skewed: there are few observations with very high debt-to-income ratios.
- The downward curve on the right side of the credit utilization and credit check plots suggests some minor misfitting for those larger values.

Having reviewed the diagnostic plots, there are two options. The first option is to, if we're not concerned about the issues observed, use this as the final model; if going this route, it is important to still note any abnormalities observed in the diagnostics. The second option is to try to improve the model, which is what we'll try to do with this particular model fit.

9.3.2 Options for improving the model fit

There are several options for improvement a model, including transforming variables, seeking out additional variables to fill model gaps, or using more advanced methods that would account

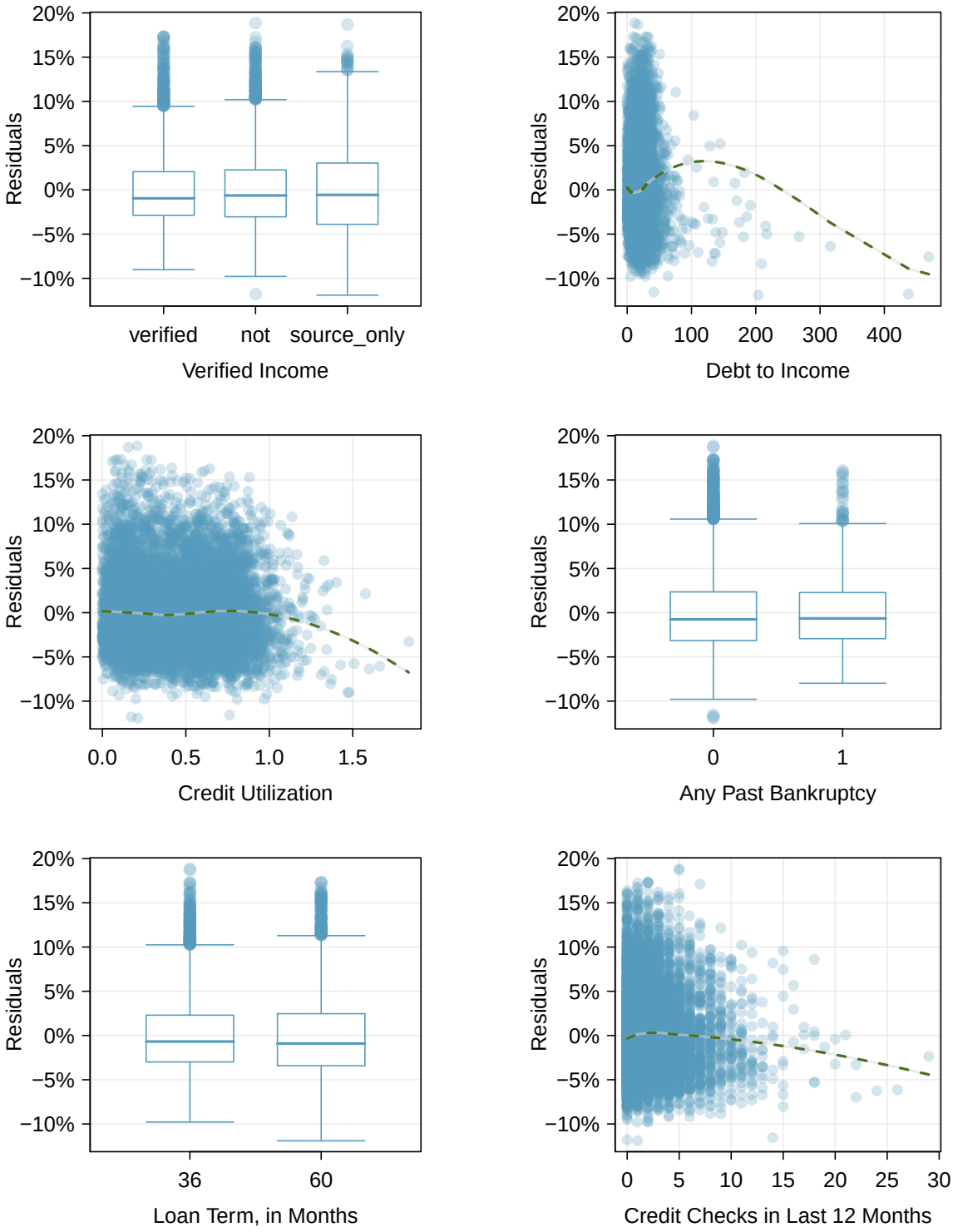


Figure 9.10: Diagnostic plots for residuals against each of the predictors. For the box plots, we're looking for notable differences in variability. For numerical predictors, we also check for trends or other structure in the data.

for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome.

The main concern for the initial model is that there is a notable nonlinear relationship between the debt-to-income variable observed in Figure 9.10. To resolve this issue, we're going to consider a couple strategies for adjusting the relationship between the predictor variable and the outcome.

Let's start by taking a look at a histogram of `debt_to_income` in Figure 9.11. The variable is extremely skewed, and upper values will have a lot of leverage on the fit. Below are several options:

- log transformation ($\log x$),
- square root transformation ($\sqrt{x}$),
- inverse transformation ($1/x$),
- truncation (cap the max value possible)

If we inspected the data more closely, we'd observe some instances where the variable takes a value of 0, and since $\log(0)$ and $1/x$ are undefined when $x = 0$, we'll exclude these transformations from further consideration.[10] A square root transformation is valid for all values the variable takes, and truncating some of the larger observations is also a valid approach. We'll consider both of these approaches.

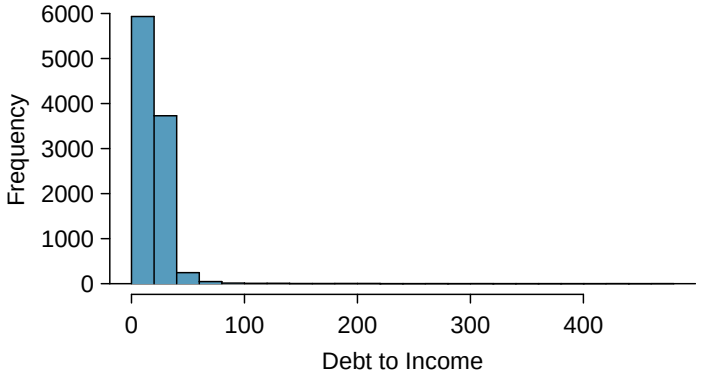


Figure 9.11: Histogram of `debt_to_income`, where extreme skew is evident.

To try transforming the variable, we make two new variables representing the transformed versions:

Square root. We create a new variable, `sqrt_debt_to_income`, where all the values are simply the square roots of the values in `debt_to_income`, and then refit the model as before. The result is shown in the left panel of Figure 9.12. The square root pulled in the higher values a bit, but the fit still doesn't look great since the smoothed line is still wavy.

Truncate at 50. We create a new variable, `debt_to_income_50`, where any values in `debt_to_income` that are greater than 50 are shrunk to exactly 50. Refitting the model once more, the diagnostic plot for this new variable is shown in the right panel of Figure 9.12. Here the fit looks much more reasonable, so this appears to be a reasonable approach.

The downside of using transformations is that it reduces the ease of interpreting the results. Fortunately, since the truncation transformation only affects a relatively small number of cases, the interpretation isn't dramatically impacted.

As a next step, we'd evaluate the new model using the truncated version of `debt_to_income`, we would complete all the same procedures as before. The other two issues noted while inspecting diagnostics in Section 9.3.1 are still present in the updated model. If we choose to report this model, we would want to also discuss these shortcomings to be transparent in our work. Depending on what the model will be used, we could either try to bring those under control, or we could stop since those issues aren't severe. Had the non-constant variance been a little more dramatic, it would be a

[10] There are ways to make them work, but we'll leave those options to a later course.

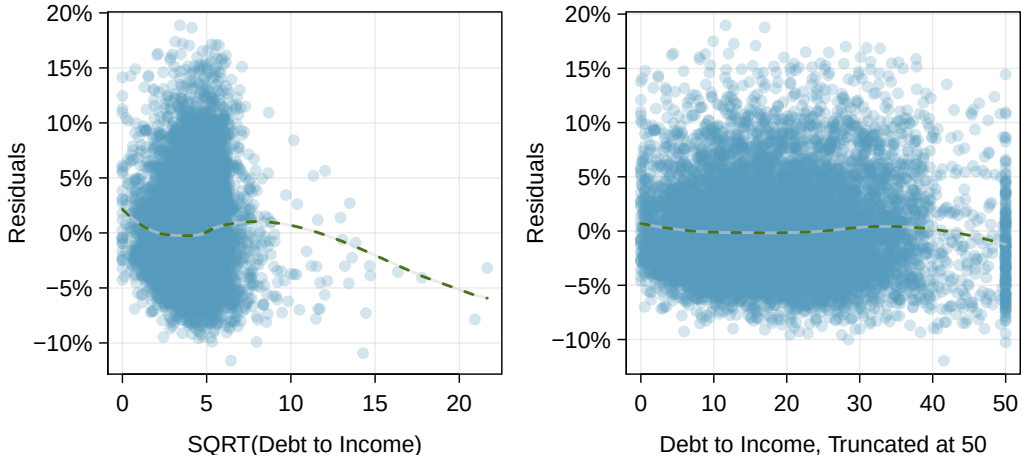


Figure 9.12: Histogram of `debt_to_income`, where extreme skew is evident.

higher priority. Ultimately we decided that the model was reasonable, and we report its final form here:

$$\begin{aligned}\widehat{rate} = \; & 1.562 + 1.002 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.436 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.048 \times \texttt{debt\_to\_income\_50} + 4.694 \times \texttt{credit\_util} + 0.394 \times \texttt{bankruptcy} \\ & + 0.153 \times \texttt{term} + 0.223 \times \texttt{credit\_check}\end{aligned}$$

A sharp eye would notice that the coefficient for `debt_to_income_50` is more than twice as large as what the coefficient had been for the `debt_to_income` variable in the earlier model. This suggests those larger values not only were points with high leverage, but they were influential points that were dramatically impacting the coefficient.

"ALL MODELS ARE WRONG, BUT SOME ARE USEFUL" -GEORGE E.P. BOX

The truth is that no model is perfect. However, even imperfect models can be useful. Reporting a flawed model can be reasonable so long as we are clear and report the model's shortcomings.

Don't report results when assumptions are grossly violated. While there is a little leeway in model assumptions, don't go too far. If model assumptions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help.

9.4 Multiple regression case study: Mario Kart

We'll consider Ebay auctions of a video game called *Mario Kart* for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables. For instance, all other characteristics held constant, are longer auctions associated with higher or lower prices? And, on average, how much more do buyers tend to pay for additional Wii wheels (plastic steering wheels that attach to the Wii controller) in auctions? Multiple regression will help us answer these and other questions.

9.4.1 Data set and the full model

The `mario_kart` data set includes results from 141 auctions. Four observations from this data set are shown in Figure 9.13, and descriptions for each variable are shown in Figure 9.14. Notice that the condition and stock photo variables are indicator variables, similar to `bankruptcy` in the `loan` data set.

	price	cond_new	stock_photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

Figure 9.13: Four observations from the `mario_kart` data set.

variable	description
`price`	Final auction price plus shipping costs, in US dollars.
`cond_new`	Indicator variable for if the game is new (`1`) or used (`0`).
`stock_photo`	Indicator variable for if the auction's main photo is a stock photo.
`duration`	The length of the auction, in days, taking values from 1 to 10.
`wheels`	The number of Wii wheels included with the auction. A *Wii wheel* is an optional steering wheel accessory that holds the Wii controller.

Figure 9.14: Variables and their descriptions for the `mario_kart` data set.

GUIDED PRACTICE 9.21

We fit a linear regression model with the game's condition as a predictor of auction price. Results of this model are summarized below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	<0.0001
cond_new	10.8996	1.2583	8.66	<0.0001
				$df = 139$

Write down the equation for the model, note whether the slope is statistically different from zero, and interpret the coefficient.[11]

[11] The equation for the line may be written as

$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

Examining the regression output in Guided Practice 9.21, we can see that the p-value for `cond_new` is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one-variable model.

Sometimes there are underlying structures or relationships between predictor variables. For instance, new games sold on Ebay tend to come with more Wii wheels, which may have led to higher prices for those auctions. We would like to fit a model that includes all potentially important variables simultaneously. This would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables.

We want to construct a model that accounts for not only the game condition, as in Guided Practice 9.21, but simultaneously accounts for three other variables:

$$\widehat{\texttt{price}} = \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} + \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels}$$

Figure 9.15 summarizes the full model. Using this output, we identify the point estimates of each coefficient.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	<0.0001
cond_new	5.1306	1.0511	4.88	<0.0001
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	<0.0001
				$df = 136$

Figure 9.15: Output for the regression model where `price` is the outcome and `cond_new`, `stock_photo`, `duration`, and `wheels` are the predictors.

GUIDED PRACTICE 9.22

Write out the model's equation using the point estimates from Figure 9.15. How many predictors are there in this model?[12]

GUIDED PRACTICE 9.23

What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?[13]

GUIDED PRACTICE 9.24

Compute the residual of the first observation in Figure 9.13 using the equation identified in Guided Practice 9.22.[14]

The `cond_new` is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. This means the 10.90 model coefficient predicts an extra $10.90 for those games that are new versus those that are used.

[12] $\widehat{price} = 36.21 + 5.13 \times \texttt{cond\_new} + 1.08 \times \texttt{stock\_photo} - 0.03 \times \texttt{duration} + 7.29 \times \texttt{wheels}$, with the $k = 4$ predictors.

[13] It is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$.

[14] $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the variables values from the observation and the equation identified in Guided Practice 9.22.

EXAMPLE 9.25

We estimated a coefficient for `cond_new` in Section 9.21 of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

(E)

If we examined the data carefully, we would see that there is collinearity among some predictors. For instance, when we estimated the connection of the outcome `price` and predictor `cond_new` using simple linear regression, we were unable to control for other variables like the number of Wii wheels included in the auction. That model was biased by the confounding variable `wheels`. When we use both variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).

9.4.2 Model selection

Let's revisit the model for the Mario Kart auction and complete model selection using backwards selection. Recall that the full model took the following form:

$$\widehat{price} = 36.21 + 5.13 \times \texttt{cond\_new} + 1.08 \times \texttt{stock\_photo} - 0.03 \times \texttt{duration} + 7.29 \times \texttt{wheels}$$

EXAMPLE 9.26

Results corresponding to the full model for the `mario_kart` data were shown in Figure 9.15 on the preceding page. For this model, we consider what would happen if dropping each of the variables in the model:

Exclude ...	`cond_new`	`stock_photo`	`duration`	`wheels`
	$R^2_{adj} = 0.6626$	$R^2_{adj} = 0.7107$	$R^2_{adj} = 0.7128$	$R^2_{adj} = 0.3487$

(E)

For the full model, $R^2_{adj} = 0.7108$. How should we proceed under the backward elimination strategy?

The third model without `duration` has the highest R^2_{adj} of 0.7128, so we compare it to R^2_{adj} for the full model. Because eliminating `duration` leads to a model with a higher R^2_{adj}, we drop `duration` from the model.

GUIDED PRACTICE 9.27

In Example 9.26, we eliminated the `duration` variable, which resulted in a model with $R^2_{adj} = 0.7128$. Let's look at if we would eliminate another variable from the model using backwards elimination:

Exclude `duration` and ...	`cond_new`	`stock_photo`	`wheels`
	$R^2_{adj} = 0.6587$	$R^2_{adj} = 0.7124$	$R^2_{adj} = 0.3414$

(G)

Should we eliminate any additional variable, and if so, which variable should we eliminate?[15]

GUIDED PRACTICE 9.28

After eliminating the auction's duration from the model, we are left with the following reduced model:

$$\widehat{price} = 36.05 + 5.18 \times \texttt{cond\_new} + 1.12 \times \texttt{stock\_photo} + 7.30 \times \texttt{wheels}$$

(G)

How much would you predict for the total price for the Mario Kart game if it was used, used a stock photo, and included two wheels and put up for auction during the time period that the Mario Kart data were collected?[16]

[15] Removing any of the three remaining variables would lead to a decrease in R^2_{adj}, so we should not remove any additional variables from the model after we removed `duration`.

G

GUIDED PRACTICE 9.29

Would you be surprised if the seller from Guided Practice 9.28 didn't get the exact price predicted?[17]

9.4.3 Checking model assumptions using graphs

Let's take a closer look at the diagnostics for the Mario Kart model to check if the model we have identified is reasonable.

Check for outliers. A histogram of the residuals is shown in Figure 9.16. With a data set well over a hundred, we're primarily looking for major outliers. While one minor outlier appears on the upper end, it is not a concern for this large of a data set.

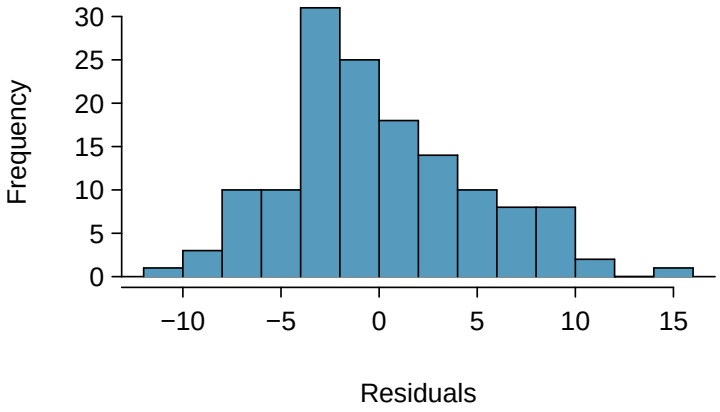


Figure 9.16: Histogram of the residuals. No clear outliers are evident.

Absolute values of residuals against fitted values. A plot of the absolute value of the residuals against their corresponding fitted values ($\hat{y}_i$) is shown in Figure 9.17. We don't see any obvious deviations from constant variance in this example.

Residuals in order of their data collection. A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 9.18. Here we see no structure that indicates a problem.

Residuals against each predictor variable. We consider a plot of the residuals against the `cond_new` variable, the residuals against the `stock_photo` variable, and the residuals against the `wheels` variable. These plots are shown in Figure 9.19. For the two-level condition variable, we are guaranteed not to see any remaining trend, and instead we are checking that the variability doesn't fluctuate across groups, which it does not. However, looking at the stock photo variable, we find that there is some difference in the variability of the residuals in the two groups. Additionally, when we consider the residuals against the `wheels` variable, we see some possible structure. There appears to be curvature in the residuals, indicating the relationship is probably not linear.

As with the `loans` analysis, we would summarize diagnostics when reporting the model results. In the case of this auction data, we would report that there appears to be non-constant variance in the stock photo variable and that there may be a nonlinear relationship between the total price and the number of wheels included for an auction. This information would be important to buyers and sellers who may review the analysis, and omitting this information could be a setback to the very people who the model might assist.

[16] We would plug in `0` for `cond_new` `1` for `stock_photo`, and `2` for `wheels` into the equation, which would return \$51.77, which is the total price we would expect for the auction.

[17] No. The model provides the *average* auction price we would expect, and the price for one auction to the next will continue to vary a bit (but less than what our prediction would be without the model).

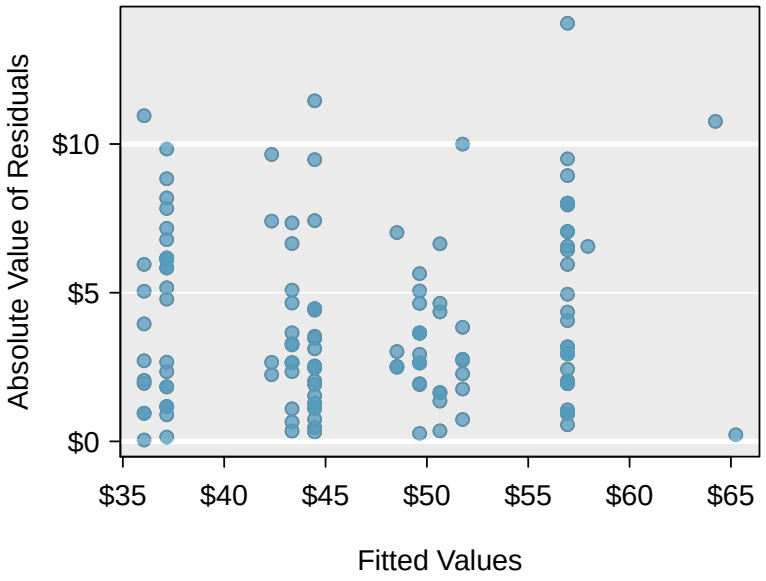


Figure 9.17: Absolute value of the residuals against the fitted values. No patterns are evident.

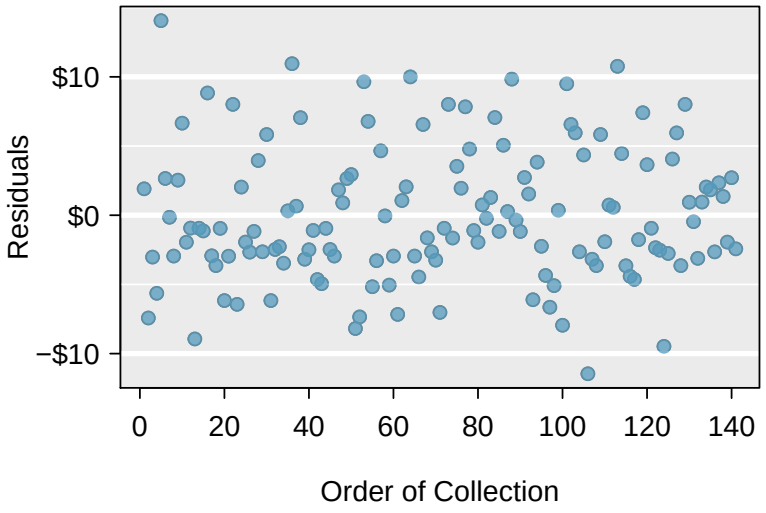


Figure 9.18: Residuals in the order that their corresponding observations were collected. There are no evident patterns.

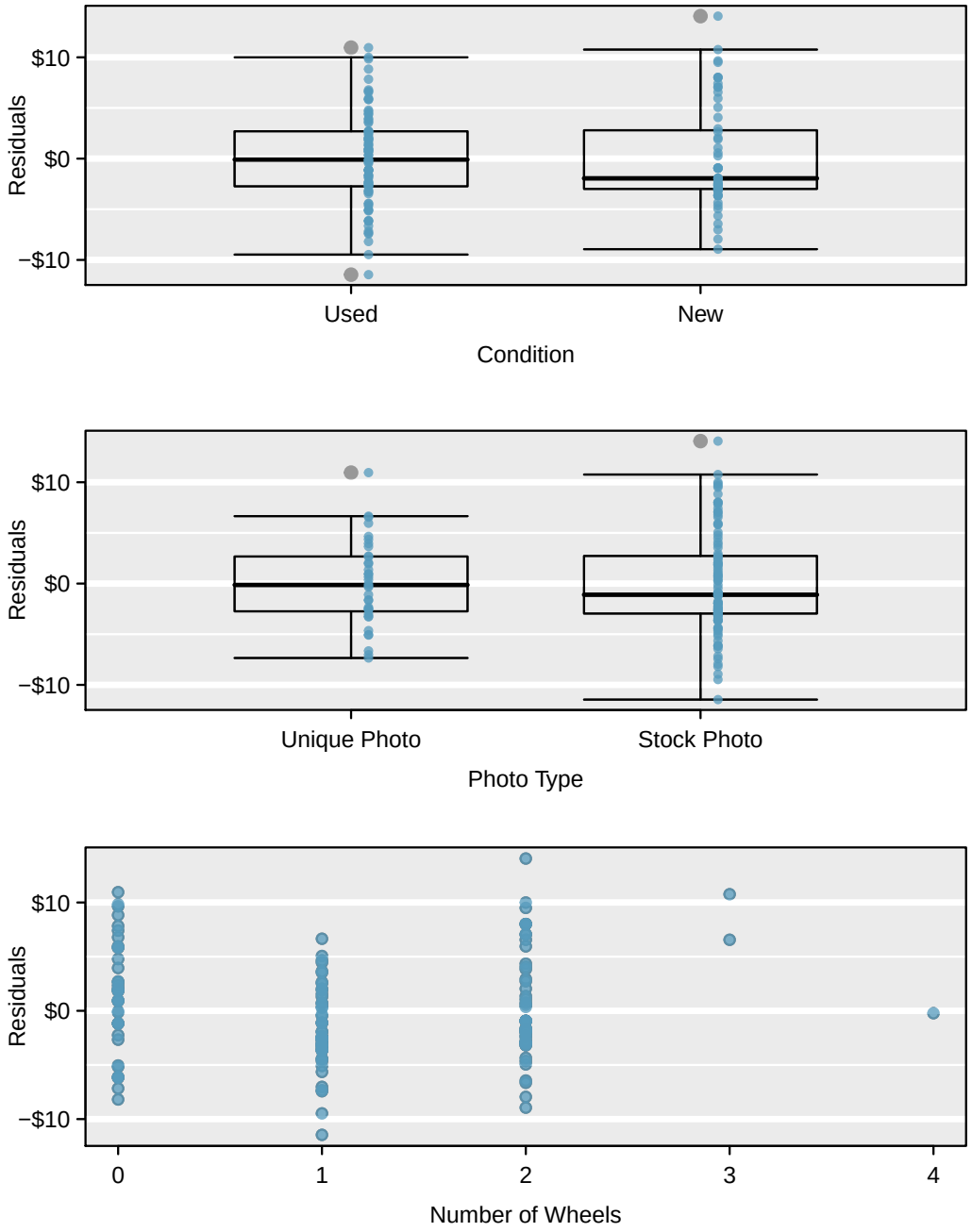


Figure 9.19: For the condition and stock photo variables, we check for differences in the distribution shape or variability of the residuals. In the case of the stock photos variable, we see a little less variability in the unique photo group than the stock photo group. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable in the bottom plot.

9.5 Introduction to logistic regression

In this section we introduce **logistic regression** as a tool for building models when there is a categorical response variable with two levels, e.g. yes and no. Logistic regression is a type of **generalized linear model** (**GLM**) for response variables where regular multiple regression does not work very well. In particular, the response variable in these settings often takes a form where residuals look completely different from the normal distribution.

GLMs can be thought of as a two-stage modeling approach. We first model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression. Ultimately, the application of a GLM will feel very similar to multiple regression, even if some of the details are different.

9.5.1 Resume data

We will consider experiment data from a study that sought to understand the effect of race and sex on job application callback rates; details of the study and a link to the data set may be found in Appendix A.9. To evaluate which factors were important, job postings were identified in Boston and Chicago for the study, and researchers created many fake resumes to send off to these jobs to see which would elicit a callback. The researchers enumerated important characteristics, such as years of experience and education details, and they used these characteristics to randomly generate the resumes. Finally, they randomly assigned a name to each resume, where the name would imply the applicant's sex and race.

The first names that were used and randomly assigned in this experiment were selected so that they would predominantly be recognized as belonging to Black or White individuals; other races were not considered in this study. While no name would definitively be inferred as pertaining to a Black individual or to a White individual, the researchers conducted a survey to check for racial association of the names; names that did not pass this survey check were excluded from usage in the experiment. You can find the full set of names that did pass the survey test and were ultimately used in the study in Figure 9.20. For example, Lakisha was a name that their survey indicated would be interpreted as a Black woman, while Greg was a name that would generally be interpreted to be associated with a White male.

first_name	race	sex
Aisha	black	female
Allison	white	female
Anne	white	female
Brad	white	male
Brendan	white	male
Brett	white	male
Carrie	white	female
Darnell	black	male
Ebony	black	female
Emily	white	female
Geoffrey	white	male
Greg	white	male

first_name	race	sex
Hakim	black	male
Jamal	black	male
Jay	white	male
Jermaine	black	male
Jill	white	female
Kareem	black	male
Keisha	black	female
Kenya	black	female
Kristen	white	female
Lakisha	black	female
Latonya	black	female
Latoya	black	female

first_name	race	sex
Laurie	white	female
Leroy	black	male
Matthew	white	male
Meredith	white	female
Neil	white	male
Rasheed	black	male
Sarah	white	female
Tamika	black	female
Tanisha	black	female
Todd	white	male
Tremayne	black	male
Tyrone	black	male

Figure 9.20: List of all 36 unique names along with the commonly inferred race and sex associated with these names.

The response variable of interest is whether or not there was a callback from the employer for the applicant, and there were 8 attributes that were randomly assigned that we'll consider, with special interest in the race and sex variables. Race and sex are **protected classes** in the United States, meaning they are not legally permitted factors for hiring or employment decisions. The full set of attributes considered is provided in Figure 9.21.

All of the attributes listed on each resume were randomly assigned. This means that no attributes that might be favorable or detrimental to employment would favor one demographic over

variable	description
`callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.
`military`	Indicator for if the resume listed any military experience.
`email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only `male` and `female` in this study), implied by the first name listed on the resume.

Figure 9.21: Descriptions for the `callback` variable along with 8 other variables in the `resume` data set. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

another on these resumes. Importantly, due to the experimental nature of this study, we can infer causation between these variables and the callback rate, if the variable is statistically significant. Our analysis will allow us to compare the practical importance of each of the variables relative to each other.

9.5.2 Modeling the probability of an event

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 (in our application, this represents a callback for the resume) with probability p_i and the value 0 with probability $1 - p_i$. Because each observation has a slightly different context, e.g. different education level or a different number of years of experience, the probability p_i will differ for each observation. Ultimately, it is this probability that we model in relation to the predictor variables: we will examine which resume characteristics correspond to higher or lower callback rates.

NOTATION FOR A LOGISTIC REGRESSION MODEL

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the resume application, Y_i will be used to represent whether resume i received a callback ($Y_i = 1$) or not ($Y_i = 0$).

The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation i, $x_{2,i}$ is the value of variable 2 for observation i, and so on.

The logistic regression model relates the probability a resume would receive a callback (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \tag{9.30}$$

We want to choose a transformation in the equation that makes practical and mathematical sense. For example, we want a transformation that makes the range of possibilities on the left hand side of the equation equal to the range of possibilities for the right hand side; if there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range. A common transformation for p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e\left(\frac{p_i}{1 - p_i}\right)$$

The logit transformation is shown in Figure 9.22. Below, we rewrite the equation relating Y_i to its

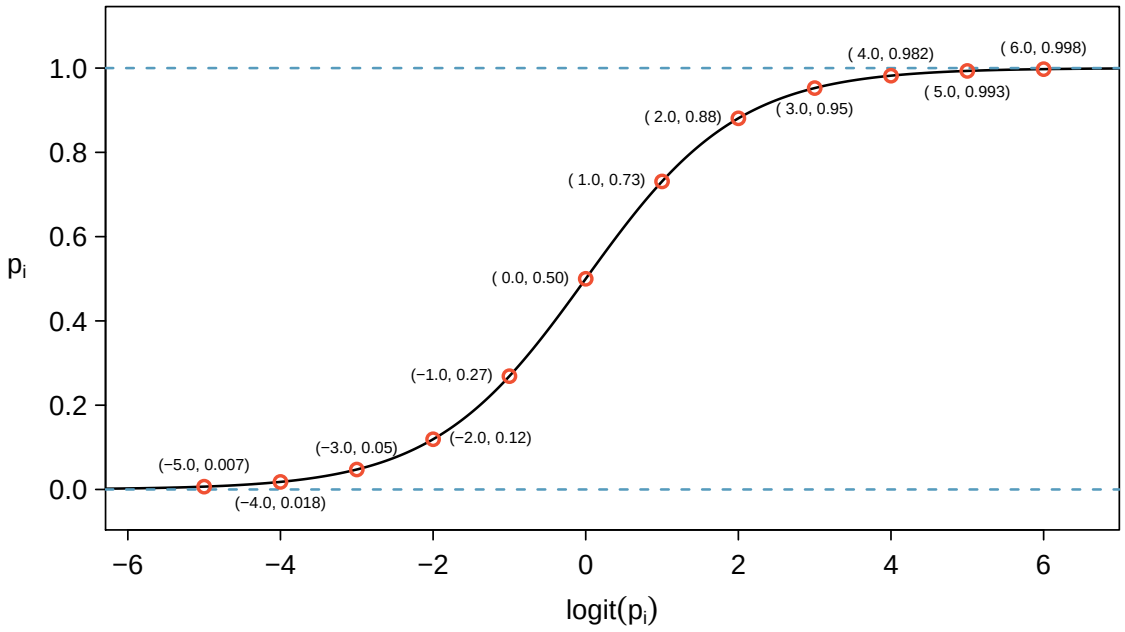


Figure 9.22: Values of p_i against values of $logit(p_i)$.

predictors using the logit transformation of p_i:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

In our resume example, there are 8 predictor variables, so $k = 8$. While the precise choice of a logit function isn't intuitive, it is based on theory that underpins generalized linear models, which is beyond the scope of this book. Fortunately, once we fit a model using software, it will start to feel like we're back in the multiple regression context, even if the interpretation of the coefficients is more complex.

EXAMPLE 9.31

We start by fitting a model with a single predictor: `honors`. This variable indicates whether the applicant had any type of honors listed on their resume, such as employee of the month. The following logistic regression model was fit using statistical software:

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.4998 + 0.8668 \times \texttt{honors}$$

(a) If a resume is randomly selected from the study and it does not have any honors listed, what is the probability resulted in a callback?

(E) (b) What would the probability be if the resume did list some honors?

(a) If a randomly chosen resume from those sent out is considered, and it does not list honors, then `honors` takes value 0 and the right side of the model equation equals -2.4998. Solving for p_i: $\frac{e^{-2.4998}}{1+e^{-2.4998}} = 0.076$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we do the same for this probability: $\hat{p}_i = 0.076$.

(b) If the resume had listed some honors, then the right side of the model equation is $-2.4998 + 0.8668 \times 1 = -1.6330$, which corresponds to a probability $\hat{p}_i = 0.163$.

Notice that we could examine -2.4998 and -1.6330 in Figure 9.22 to estimate the probability before formally calculating the value.

To convert from values on the logistic regression scale (e.g. -2.4998 and -1.6330 in Exam-

ple 9.31), use the following formula, which is the result of solving for p_i in the regression model:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 \ + \ e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute the point estimates for the parameters (the β_i) so that we can make use of this formula. In Example 9.31, the probabilities were calculated as

$$\frac{e^{-2.4998}}{1 \ + \ e^{-2.4998}} = 0.076 \qquad\qquad \frac{e^{-2.4998+0.8668}}{1 \ + \ e^{-2.4998+0.8668}} = 0.163$$

While knowing whether a resume listed honors provides some signal when predicting whether or not the employer would call, we would like to account for many different variables at once to understand how each of the different resume characteristics affected the chance of a callback.

9.5.3 Building the logistic model with many variables

We used statistical software to fit the logistic regression model with all 8 predictors described in Figure 9.21. Like multiple regression, the result may be presented in a summary table, which is shown in Figure 9.23. The structure of this table is almost identical to that of multiple regression; the only notable difference is that the p-values are calculated using the normal distribution rather than the t-distribution.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.6632	0.1820	-14.64	<0.0001
job_city: *Chicago*	-0.4403	0.1142	-3.85	0.0001
college_degree	-0.0666	0.1211	-0.55	0.5821
years_experience	0.0200	0.0102	1.96	0.0503
honors	0.7694	0.1858	4.14	<0.0001
military	-0.3422	0.2157	-1.59	0.1127
email_address	0.2183	0.1133	1.93	0.0541
race: *white*	0.4424	0.1080	4.10	<0.0001
sex: *male*	-0.1818	0.1376	-1.32	0.1863

Figure 9.23: Summary table for the full logistic regression model for the resume callback example.

Just like multiple regression, we could trim some variables from the model. Here we'll use a statistic called **Akaike information criterion (AIC)**, which is an analog to how we used adjusted R-squared in multiple regression, and we look for models with a lower AIC through a backward elimination strategy. After using this criteria, the `college_degree` variable is eliminated, giving the smaller model summarized in Figure 9.24, which is what we'll rely on for the remainder of this section.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7162	0.1551	-17.51	<0.0001
job_city: *Chicago*	-0.4364	0.1141	-3.83	0.0001
years_experience	0.0206	0.0102	2.02	0.0430
honors	0.7634	0.1852	4.12	<0.0001
military	-0.3443	0.2157	-1.60	0.1105
email_address	0.2221	0.1130	1.97	0.0494
race: *white*	0.4429	0.1080	4.10	<0.0001
sex: *male*	-0.1959	0.1352	-1.45	0.1473

Figure 9.24: Summary table for the logistic regression model for the resume callback example, where variable selection has been performed using AIC.

EXAMPLE 9.32

The `race` variable had taken only two levels: `black` and `white`. Based on the model results, was race a meaningful factor for if a prospective employer would call back?

(E) We see that the p-value for this coefficient is very small (very nearly zero), which implies that race played a statistically significant role in whether a candidate received a callback. Additionally, we see that the coefficient shown corresponds to the level of `white`, and it is positive. This positive coefficient reflects a positive gain in callback rate for resumes where the candidate's first name implied they were White. The data provide very strong evidence of racism by prospective employers that favors resumes where the first name is typically interpreted to be White.

The coefficient of `racewhite` in the full model in Figure 9.23, is nearly identical to the model shown in Figure 9.24. The predictors in this experiment were thoughtfully laid out so that the coefficient estimates would typically not be much influenced by which other predictors were in the model, which aligned with the motivation of the study to tease out which effects were important to getting a callback. In most observational data, it's common for point estimates to change a little, and sometimes a lot, depending on which other variables are included in the model. Collinearity can also occur in experiments, but in this case the experiment was designed in such a way that collinearity it was not an issue.

EXAMPLE 9.33

Use the model summarized in Figure 9.24 to estimate the probability of receiving a callback for a job in Chicago where the candidate lists 14 years experience, no honors, no military experience, includes an email address, and has a first name that implies they are a White male.

We can start by writing out the equation using indicator functions and coefficients from the model, then we can add in the corresponding values of each variable for this individual:

(E)
$$\begin{aligned} log\left(\frac{p}{1-p}\right) &= -2.7162 - 0.4364 \times 1_{Chicago} + 0.0206 \times (\text{years experience}) + 0.7634 \times 1_{honors} \\ &\quad - 0.3443 \times 1_{military} + 0.2221 \times 1_{email} + 0.4429 \times 1_{white} - 0.1959 \times 1_{male} \\ &= -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 + 0.7634 \times 0 \\ &\quad - 0.3443 \times 0 + 0.2221 \times 1 + 0.4429 \times 1 - 0.1959 \times 1 \\ &= -2.3955 \end{aligned}$$

We can now back-solve for p: the chance such an individual will receive a callback is about 8.35%.

EXAMPLE 9.34

Compute the probability of a callback for an individual with a name commonly inferred to be from a Black male but who otherwise has the same characteristics as the one described in Example 9.33.

(E) We can complete the same steps for an individual with the same characteristics who is Black, where the only difference in the calculation is that the indicator variable 1_{white} will take a value of `0`. Doing so yields a probability of 0.0553. Let's compare the results with those of Example 9.33.

In practical terms, an individual perceived as White based on their first name would need to apply to $\frac{1}{0.0835} \approx 12$ jobs on average to receive a callback, while an individual perceived as Black based on their first name would need to apply to $\frac{1}{0.0553} \approx 18$ jobs on average to receive a callback. That is, applicants who are perceived as Black need to apply to 50% more employers to receive a callback than someone who is perceived as White based on their first name for jobs like those in the study.

What we've quantified in this section is alarming and disturbing. However, one aspect that makes this racism so difficult to address is that the experiment, as well-designed as it is, cannot send us much signal about which employers are discriminating. It is only possible to say that discrimination is happening, even if we cannot say which particular callbacks – or non-callbacks – represent discrimination. Finding strong evidence of racism for individual cases is a persistent

challenge in enforcing anti-discrimination laws.

9.5.4 Diagnostics for the callback rate model

LOGISTIC REGRESSION CONDITIONS

There are two key conditions for fitting a logistic regression model:

1. Each outcome Y_i is independent of the other outcomes.
2. Each predictor x_i is linearly related to logit(p_i) if all other predictors are held constant.

The first logistic regression model condition – independence of the outcomes – is reasonable for the experiment since characteristics of resumes were randomly assigned to the resumes that were sent out.

The second condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 4870 resume submissions in the data set! Let's first visualize these data by plotting the true classification of the resumes against the model's fitted probabilities, as shown in Figure 9.25.

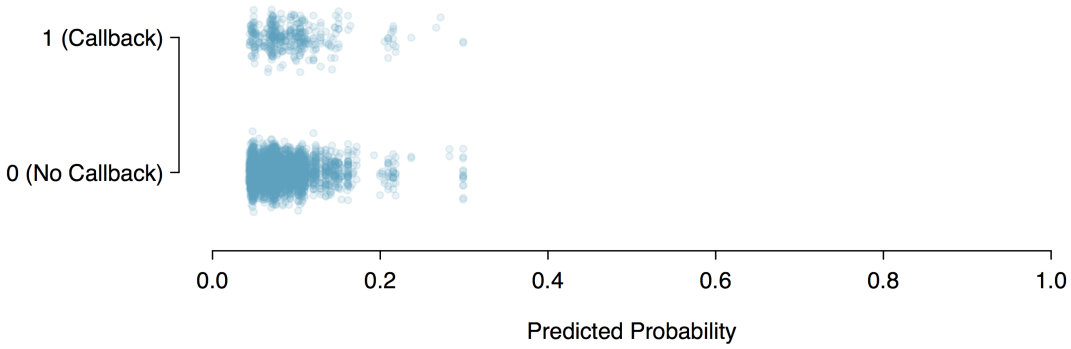


Figure 9.25: The predicted probability that each of the 4870 resumes results in a callback. Noise (small, random vertical shifts) have been added to each point so points with nearly identical values aren't plotted exactly on top of one another.

We'd like to assess the quality of the model. For example, we might ask: if we look at resumes that we modeled as having a 10% chance of getting a callback, do we find about 10% of them actually receive a callback? We can check this for groups of the data by constructing a plot as follows:

1. Bucket the data into groups based on their predicted probabilities.
2. Compute the average predicted probability for each group.
3. Compute the observed probability for each group, along with a 95% confidence interval.
4. Plot the observed probabilities (with 95% confidence intervals) against the average predicted probabilities for each group.

The points plotted should fall close to the line $y = x$, since the predicted probabilities should be similar to the observed probabilities. We can use the confidence intervals to roughly gauge whether anything might be amiss. Such a plot is shown in Figure 9.26.

Additional diagnostics may be created that are similar to those featured in Section 9.3. For instance, we could compute residuals as the observed outcome minus the expected outcome ($e_i = Y_i - \hat{p}_i$), and then we could create plots of these residuals against each predictor. We might also create a plot like that in Figure 9.26 to better understand the deviations.

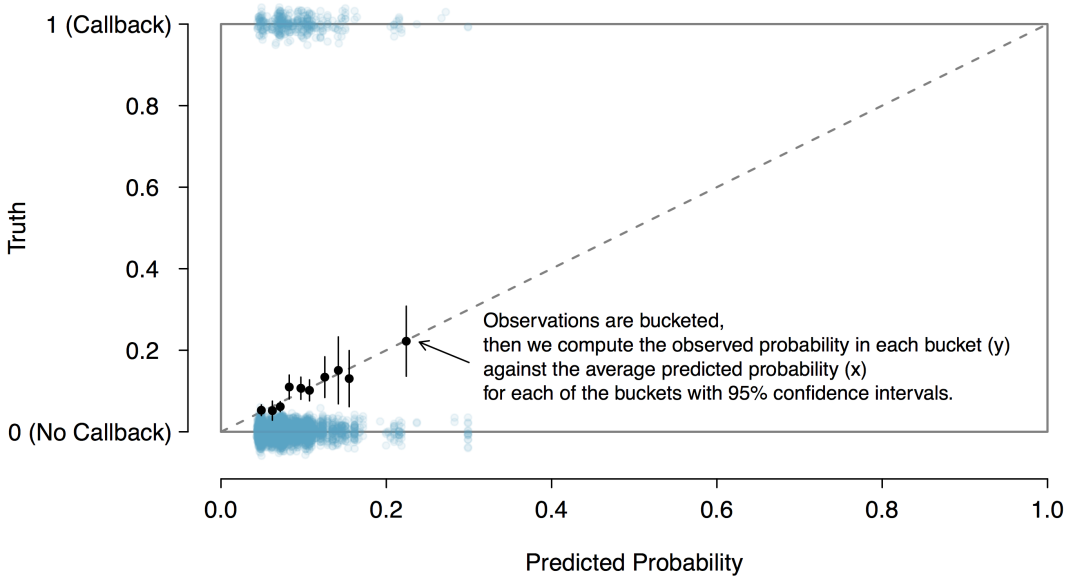


Figure 9.26: The dashed line is within the confidence bound of the 95% confidence intervals of each of the buckets, suggesting the logistic fit is reasonable.

9.5.5 The mathematics behind discrimination

Any form of discrimination is concerning, and this is why we decided it was so important to discuss this topic using data. This study also only examined discrimination in a single aspect: whether a prospective employer would call a candidate who submitted their resume. There was a 50% higher barrier for resumes simply when the candidate had a first name that was perceived to be from a Black individual. It's unlikely that discrimination would stop there.

Of course, discrimination can happen to anyone. Yet, discrimination against non-oppressed groups is considered to be much less impactful than the discrimination experienced by minorities. *Why?*

EXAMPLE 9.35

Let's consider a sex-imbalanced company that consists of 20% women and 80% men,[18] and we'll suppose that the company is very large, consisting of perhaps 20,000 employees. Suppose when someone goes up for promotion at this company, 5 of their colleagues are randomly chosen to provide feedback on their work.

Now let's imagine that 10% of the people in the company are prejudiced against the other sex. That is, 10% of men are prejudiced against women, and similarly, 10% of women are prejudiced against men.

Who is discriminated against more at the company, men or women?

(E)

Let's suppose we took 100 men who have gone up for promotion in the past few years. For these men, $5 \times 100 = 500$ random colleagues will be tapped for their feedback, of which, about 20% will be women (100 women). Of these 100 women, 10 are expected to be biased against the man they are reviewing. Then, of the 500 colleagues reviewing them, men will experience discrimination by about 2% of their colleagues when they go up for promotion.

Let's do a similar calculation for 100 women who have gone up for promotion in the last few years. They will also have 500 random colleagues providing feedback, of which about 400 (80%) will be men. Of these 400 men, about 40 (10%) hold a bias against women. Of the 500 colleagues providing feedback on the promotion packet for these women, 8% of the colleagues hold a bias against the women.

Example 9.35 highlights something profound: even if each demographic has the same degree of prejudice against the other demographic, minority groups can experience those effects more frequently. Additionally, if we would complete a handful of examples like the one above with different numbers, we'd learn that the greater the imbalance in the population groups, the more the minority is disproportionately impacted. For example, if a proportion p of a company are women and the rest of the company consists rest men, then the ratio of rates of discrimination against women vs men would be given by $\frac{1-p}{p}$; this ratio is always greater than 1 when $p < 0.5$. That is, whenever individuals are equally likely to discriminate, the minority group would always encounter more discrimination, and the degree of that discrimination will be larger the greater the imbalance in the population.

One considerable omission from the company example above is that discrimination can also happen for people from oppressed groups to discriminate against others within their own oppressed group. For an interesting study on this topic, please see

> *Milkman KL, Akinola M, Chugh D. 2015. What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway Into Organizations. Journal of Applied Psychology, 100:6, p1678-1712.*

The paper's abstract summarizes the findings, and substantial detail of the analysis is provided within the paper. We've also made the data set available, which is noted in Appendix A.9 so that you may also explore it directly. [If we do not obtain the data from this study, then need to delete the last sentence.]

No discrimination has a place in our society, be it discrimination against a minority group or a majority group. Yet we cannot deny the mathematics behind discrimination: minority groups are much more prone to the negative impacts from discrimination than majority groups. While we will stand up to discrimination in all its forms, we recognize that the biggest problems we face today is discrimination against minorities and oppressed groups.

We close this book on this serious topic, and we hope it inspires you to think about the power of reasoning with data. Whether it is with a formal statistical model or by using critical thinking skills to structure a problem, we hope the ideas you have learned will help you do more and do better in life.

[18] A more thoughtful example would include non-binary individuals.

Appendix A

Data sets within the text

Each data set within the text is described in this appendix, and there is a corresponding page for each of these data sets at **openintro.org/data**. This page also includes additional data sets that can be used for honing your skills. Each data set has its own page with the following information:

- Description of each data set.
- Detailed overview of each data set's variables.
- CSV download.
- R object file download.

Over time we will also expand the information available on these pages.
[Redirects must be created for each link in this appendix.]

A.1 Introduction to data

1.1 [`stent30`, `stent365`] The stent data is split across two data sets, one for the 0-30 day and one for the 0-365 day results.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
NY Times article: www.nytimes.com/2011/09/08/health/research/08stent.html.

1.2 [`loan50`, `loans_full_schema`] This data comes from Lending Club (lendingclub.com), which provides a large set of data on the people who received loans through their platform. The data used in the textbook comes from a sample of the loans made in Q1 (Jan, Feb, March) 2018.

1.2 [`county`, `county_complete`] These data come from several government sources. For those variables included in the county data set, only the most recent data is reported, as of what was available in late 2018. Data prior to 2011 is all from census.gov, where the specific Quick Facts page providing the data is no longer available. The more recent data comes from USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), and American Community Survey (census.gov/programs-surveys/acs).

1.3 No data sets were described in this section.

1.3 The Nurses' Health Study was mentioned. For more information on this data set, see www.channing.harvard.edu/nhs

1.4 The study we had in mind when discussing the simple randomization (no blocking) study was Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

A.2 Summarizing data

2.1 [loan50, county] These data sets are described in Data Appendix A.1.

2.2 [loan50, county] These data sets are described in Data Appendix A.1.

2.3 [malaria] Lyke et al. 2017. PfSPZ vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. PNAS 114(10):2711-2716. http://www.pnas.org/content/114/10/2711

A.3 Probability

3.1 [loan50, county] These data sets are described in Data Appendix A.1.

3.1 [playing_cards] A table describing the 52 cards in a standard deck.

3.2 [family_college] A simulated data set based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

3.2 [smallpox] Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

3.2 [Mammogram screening, probabilities.] The probabilities reported were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

3.2 [Jose campus visits, probabilities, no data link] This example was entirely made up.

3.3 No data sets were described in this section.

3.4 [Course material purchases, probabilities, no data link] The probabilities for course materials purchases were created as hypotheticals. [Should this even be mentioned?]

3.4 [Auctions for TV and toaster, no data link] The statistics used were created as hypotheticals. [Should this even be mentioned?]

3.4 [stocks_18] Monthly returns for Caterpillar, Exxon Mobil Corp, and Google for November 2015 to October 2018.

3.5 [fcid] This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

A.4 Distributions of random variables

4.1 [SAT and ACT score distributions] The SAT score data comes from the 2018 distribution, which is provided at
reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf
The ACT score data is available at
act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf
We also acknowledge that the actual ACT score distribution is *not* nearly normal. However, since the topic is very accessible, we decided to keep the context and examples.

4.1 [Male heights] The distribution is based on the USDA Food Commodity Intake Database.

4.1 [possum] The distribution parameters are based on a sample of possums from Australia and New Guinea. The original source of this data is as follows. Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala).* Australian Journal of Zoology 43: 449-458.

4.2 [Exceeding insurance deductible] These statistics were made up but are possible values one might observe for low-deductible plans.

4.3 [Exceeding insurance deductible] These statistics were made up but are possible values one might observe for low-deductible plans.

4.3 [Smoking friends] Unfortunately, we don't currently have additional information on the source for the 30% statistic, so don't consider this one as fact since we cannot verify it was from a reputable source.

4.3 [US smoking rate] The 15% smoking rate in the US figure is close to the value from the Centers for Disease Control and Prevention website, which reports a value of 14% as of the 2017 estimate:
cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

4.4 [Football kicker] This example was made up.

4.4 [Heart attack admissions] This example was made up, though the heart attack admissions are realistic for some hospitals.

4.5 [ami_occurrences] This is a simulated data set but resembles actual AMI data for New York City based on typical AMI incidence rates.

A.5 Foundations for inference

5.1 [pew_energy_2018] The actual data has more observations than were referenced in this chapter. That is, we used a subsample since it helped smooth some of the examples to have a bit more variability. The **pew_energy_2018** data set represents the full data set for each of the different energy source questions, which covers solar, wind, offshore drilling, hydrolic fracturing, and nuclear energy. The statistics used to construct the data are from the following page:

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

5.2 [pew_energy_2018] See the details for this data set above in the Section 5.1 data section.

5.2 [ebola_survey] In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014. Poll ID NY141026 on maristpoll.marist.edu.

5.3 [pew_energy_2018] See the details for this data set above in the Section 5.1 data section.

5.3 [Rosling questions] We noted much smaller samples than the Roslings' describe in their book, Factfulness, The samples we describe are similar but not the same as the actual rates. The approximate rates for the correct answers for the two questions for (sometimes different) populations discussed in the book, as reported in Factfulness, are

- 80% of the world's 1 year olds have been vaccinated against some disease: 13% get this correct (17% in the US). gapm.io/q9
- Number of children in the world in 2100: 9% correct. gapm.io/q5

Here are a few more questions and a rough percent of people who get them correct:

- In all low-income countries across the world today, how many girls finish primary school: 20%, 40%, or 60%? Answer: 60%. About 7% of people get this question correct. gapm.io/q1
- What is the life expectancy of the world today: 50 years, 60 years, or 70 years? Answer: 70 years. In the US, about 43% of people get this question correct. gapm.io/q4
- In 1996, tigers, giant pandas, and black rhinos were all listed as endangered. How many of these three species are more critically endangered today: two of them, one of them, none of them? Answer: none of them. About 7% of people get this question correct. gapm.io/q11

– How many people in the world have some access to electricity? 20%, 50%, 80%. Answer: 80%. About 22% of people get this correct. gapm.io/q12

For more information, check out the book, Factfulness.

5.3 [nuclear_survey] A simple random sample of 1,028 US adults in March 2013 found that 56% of US adults support nuclear arms reduction. www.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx

5.3 [stent30, stent365] This data is described in Data Appendix A.1.

A.6 Inference for categorical data

6.1 [Payday loans] The statistics come from the following source: pewtrusts.org/-/media/assets/2017/04/payday-loan-customers-want-more-protections-methodology.pdf

6.1 [Tire factory] The statistics were created for the purposes of performing sample size calculations.

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

6.2 [fish_oil_18] Manson JE, et al. 2018. Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer. NEJMoa1811403.

6.2 [mammogram] Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

6.2 [drone_blades] The quality control data set for quadcopter drone blades is a made-up data set for an example. We provide the simulated data in the `drone_blades` data set.

6.3 [jury] The jury data set for examining discrimination is a made-up data set an example. We provide the simulated data in the `jury` data set.

6.3 [sp500_1950_2018] Data is sourced from finance.yahoo.com.

6.4 [ask] Minson JA, Ruedy NE, Schweitzer ME. *There is such a thing as a stupid question: Question disclosure in strategic communication.* opim.wharton.upenn.edu/DPlab/papers/workingPapers/Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

6.4 [diabetes2] Zeitler P, et al. 2012. A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes. N Engl J Med.

A.7 Inference for numerical data

7.1 [Risso's dolphins] Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins.

7.1 [Croaker white fish] www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

7.1 [run17] www.cherryblossom.org

7.2 [textbooks, ucla_textbooks_f18] Data were collected by OpenIntro staff in 2010 and again in 2018. For the 2018 sample, we sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. The websites where information was retrieved: sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, and amazon.com.

7.3 [stem_cells] Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012.

7.3 [ncbirths] Birth records released by North Carolina in 2004. Unfortunately, we don't currently have additional information on the source for this data set.

7.3 [Exam versions] This is a made-up data set for an example. There isn't any explicit data set to analyze, only summary statistics.

7.4 [Blood pressure statistics] The blood pressure standard deviation for patients with blood pressure ranging from from 140 to 180 mmHg is guessed and may be a little (but likely not dramatically) imprecise from what we'd observe in actual data.

7.5 [toy_anova] Data used for Figure 7.20, which was fake data.

7.5 [mlb_players_18] Data were retrieved from mlb.mlb.com/stats. Only players with at least 100 at bats were considered during the analysis.

7.5 [class_data] These data are simulated. [Data set is currently called `classData` in the package.]

A.8 Introduction to linear regression

8.1 [simulated_scatter] Fake data used for the first three plots. The perfect linear plot uses group 4 data, where `group` variable in the data set (Figure 8.1). The group of 3 imperfect linear plots use groups 1-3 (Figure 8.2). The sinusoidal curve uses group 5 data (Figure 8.3). The curved plot of data with the curved band uses group 30 data (right panel of Figure **??**). The group of 3 scatterplots with residual plots use groups 6-8 (Figure 8.8). The correlation plots uses groups 9-19 data (Figures 8.9 and 8.10).

8.1 [possum] This data is described in Data Appendix A.4.

8.2 [elmhurst] These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

8.2 [simulated_scatter] The plots for things that can go wrong uses groups 20-23 (Figure 8.12).

8.2 [mario_kart] [Confirm data set name change.] Auction data from Ebay (ebay.com) for the game Mario Kart for the Nintendo Wii. This data set was collected in early October, 2009.

8.3 [simulated_scatter] The plots for types of outliers uses groups 24-29 (Figure 8.18).

8.4 [midterms_house] Data was retrieved from Wikipedia.

A.9 Multiple and logistic regression

9.1 [loans_full_schema] This data is described in Data Appendix A.1.

9.2 [loans_full_schema] This data is described in Data Appendix A.1.

9.3 [loans_full_schema] This data is described in Data Appendix A.1.

9.4 [mario_kart] This data is describe in Data Appendix A.8.

9.5 [resume] Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873

We did omit discussion of some structure in the data for the analysis presented: the experiment design included blocking, where typically four resumes were sent to each job: one for each inferred race/sex combination (as inferred based on the first name). We did not worry

about this blocking aspect, since accounting for the blocking would *reduce* the standard error without notably changing the point estimates for the `race` and `sex` variables versus the analysis performed in the section. That is, the most interesting conclusions in the study would be unaffected even with a more sophisticated analysis.

9.5 [`research_reply`] Milkman KL, Akinola M, Chugh D. 2015. What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway Into Organizations. Journal of Applied Psychology, 100:6, p1678-1712.

This study highlights results where fictional students contacted faculty members. The outcome of interest was whether the faculty member would reply, and the variables of interest were the race and sex of the prospective student as well as demographics of the faculty member who received the message. The authors have made the data set publicly available, and we've put it into a CSV file that is friendly for downloading through the `research_reply` data set. [Either get this data set in a sharable form or remove this reference.]

Appendix B

Distribution tables

B.1 Normal Probability Table

A **normal probability table** may be used to find percentiles of a normal distribution using a Z-score, or vice-versa. Such a table lists Z-scores and the corresponding percentiles. An abbreviated probability table is provided in Figure B.1 that we'll use for the examples in this appendix. A full table may be found on page 285.

	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure B.1: A section of the normal probability table. The percentile for a normal random variable with $Z = 1.00$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

When using a normal probability table to find a percentile for Z (rounded to two decimals), identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation. For instance, the percentile of $Z = 0.45$ is shown in row 0.4 and column 0.05 in Figure B.1: 0.6736, or the 67.36^{th} percentile.

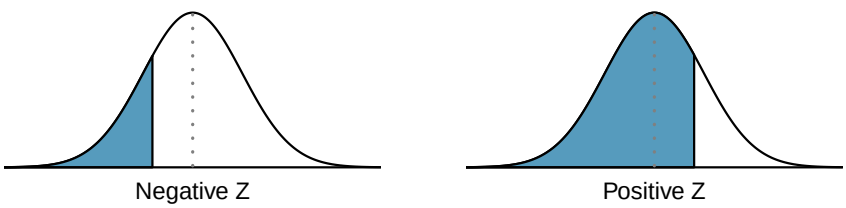
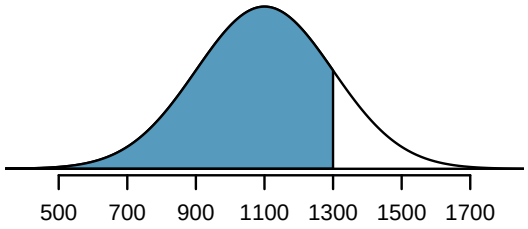


Figure B.2: The area to the left of Z represents the percentile of the observation.

EXAMPLE B.1

SAT scores follow a normal distribution, $N(1100, 200)$. Ann earned a score of 1300 on her SAT with a corresponding Z-score of $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than her. We shade the area representing those individuals in the following graph:



The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in the graph. We find this area by looking in row 1.0 and column 0.00 in the normal probability table: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

EXAMPLE B.2

How do we find an upper tail area?

The normal probability table *always* gives the area to the left. This means that if we want the area to the right, we first find the lower tail and then subtract it from 1. For instance, 84.13% of SAT takers scored below Ann, which means 15.87% of test takers scored higher than Ann.

We can also find the Z-score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z-score for the 80^{th} percentile by combining the row and column Z values: 0.84.

EXAMPLE B.3

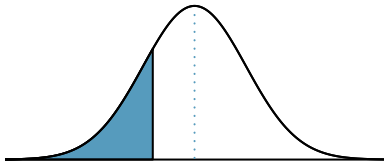
Find the SAT score for the 80^{th} percentile.

We look for the are to the value in the table closest to 0.8000. The closest value is 0.7995, which corresponds to $Z = 0.84$, where 0.8 comes from the row value and 0.04 comes from the column value. Next, we set up the equation for the Z-score and the unknown value x as follows, and then we solve for x:

$$Z = 0.84 = \frac{x - 1100}{200} \quad \rightarrow \quad x = 1268$$

The College Board scales scores to increments of 10, so the 80^{th} percentile is 1270. (Reporting 1268 would have been perfectly okay for our purposes.)

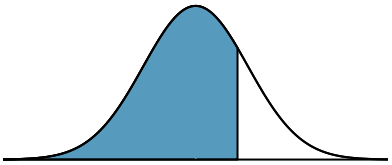
For additional details about working with the normal distribution and the normal probability table, see Section 4.1, which starts on page 99.



Negative Z

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	−0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	−0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	−0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	−0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.



	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

A **t-probability table** may be used to find tail areas of a t-distribution using a T-score, or vice-versa. Such a table lists T-scores and the corresponding percentiles. A partial **t-table** is shown in Figure B.3, and the complete table starts on page 289. Each row in the t-table represents a t-distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Figure B.3. If we want the value in this row that identifies the T-score (cutoff) for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

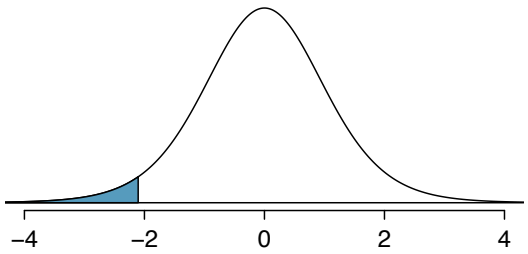
one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Figure B.3: An abbreviated look at the t-table. Each row represents a different t-distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

EXAMPLE B.4

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture and shade the area below -2.10:



To find this area, we first identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. That is, 2.5% of the distribution falls below -2.10.

In the next example we encounter a case where the exact T-score is not listed in the table.

EXAMPLE B.5

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure B.4. Estimate the proportion of the distribution falling above 1.65.

(E)

We identify the row in the t-table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

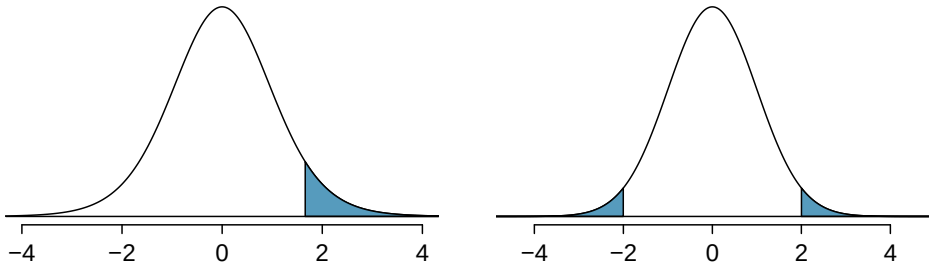


Figure B.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 475 degrees of freedom, with the area further than 2 units from 0 shaded.

EXAMPLE B.6

A t-distribution with 475 degrees of freedom is shown in the right panel of Figure B.4. Estimate the proportion of the distribution falling more than 2 units from the mean (above or below).

(E)

As before, first identify the appropriate row: $df = 475$. This row does not exist! When this happens, we use the next smaller row, which in this case is $df = 400$. Next, find the columns that capture 2.00; because $1.97 < 3 < 2.34$, we use the third and fourth columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.02 and 0.05. We use the two tail values because we are looking for two symmetric tails in the t-distribution.

(G)

GUIDED PRACTICE B.7

What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units?[1]

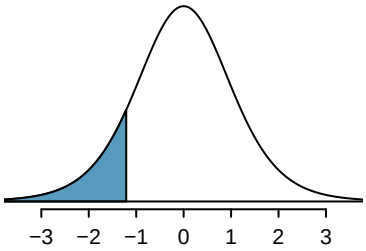
EXAMPLE B.8

Find the value of $t^\star_{18}$ using the t-table, where $t^\star_{18}$ is the cutoff for the t-distribution with 18 degrees of freedom where 95% of the distribution lies between $-t^\star_{18}$ and $+t^\star_{18}$.

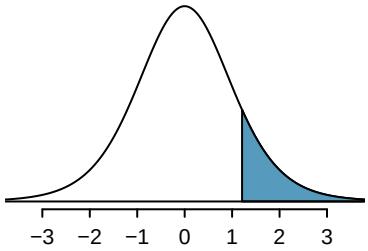
(E)

For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t-distribution is between $-t^\star_{18}$ and $t^\star_{18}$; this is the same as where the two tails have a total area of 0.05. We look in the t-table on page 287, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

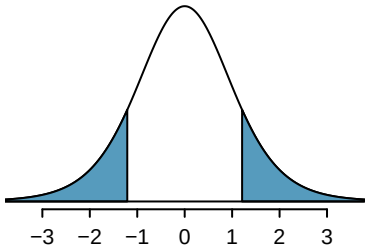
[1] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.



One Tail

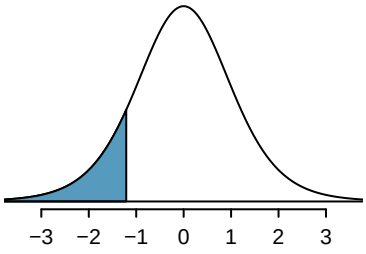


One Tail

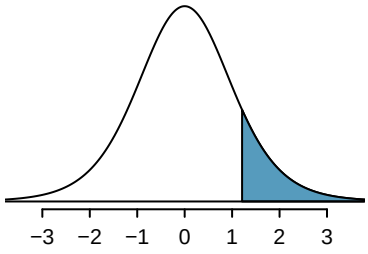


Two Tails

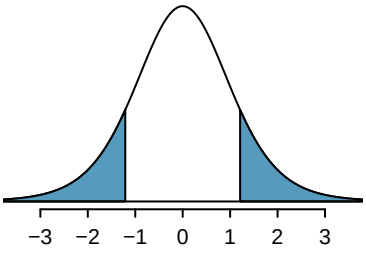
one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	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



One Tail



Two Tails

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

A **chi-square probability table** may be used to find tail areas of a chi-square distribution. The **chi-square table** is partially shown in Figure B.5, and the complete table may be found on page 292. When using a chi-square table, we examine a particular row for distributions with different degrees of freedom, and we identify a range for the area (e.g. 0.025 to 0.05). Note that the chi-square table provides upper tail values, which is different than the normal and t-distribution tables.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Figure B.5: A section of the chi-square table. A complete table is in Appendix B.3.

EXAMPLE B.9

Figure B.6(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Figure B.5 to estimate the shaded area.

E This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure B.6(a) has area 0.1.

This example was unusual, in that we observed the *exact* value in the table. In the next examples, we encounter situations where we cannot precisely estimate the tail area and must instead provide a range of values.

EXAMPLE B.10

Figure B.6(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The area above value 4.3 has been shaded; find this tail area.

E The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure B.6(b) is between 0.1 and 0.2.

EXAMPLE B.11

Figure B.6(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

E Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than* 0.3.

EXAMPLE B.12

Figure B.6(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.

E The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.

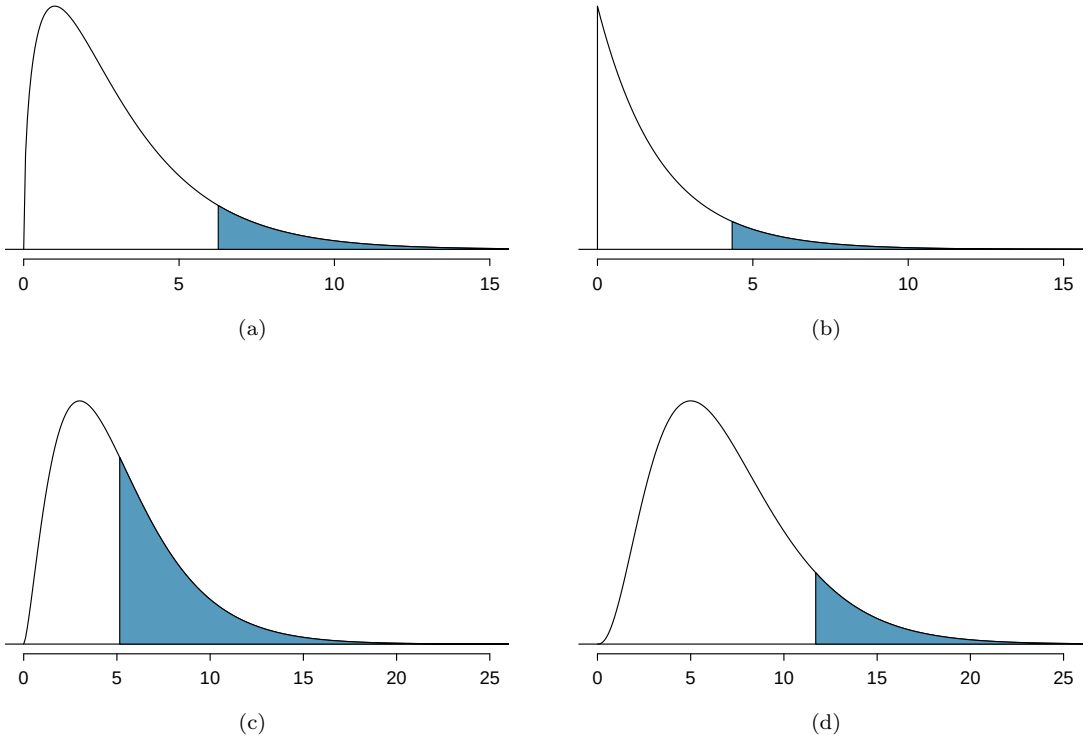


Figure B.6: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66

Index