Introduction to Probability and Statistics

Customized Course Text for STAT*2060: Statistics for Business Decisions

Derivative of
OpenIntro Statistics (Second Edition)

Original Authors

David M Diez
Quantitative Analyst
Google/YouTube

Christopher D Barr
Assistant Professor
Department of Biostatistics
Harvard School of Public Health

Mine Çetinkaya-Rundel
Assistant Professor of the Practice
Department of Statistics
Duke University

Contributing Author

Nishan Mudalige
Instructor in Statistics
Department of Mathematics and Statistics
University of Guelph

Contents

Preface

This book was based on OpenInto Statistics (Second Edition). A copy of OpenInto Statistics (Second Edition) may be downloaded as a free PDF at **openintro.org**.

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

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

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

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

The chapters of this book are as follows:

1. **Overview.** General concept of inherent certainty and the power of prediction.
1. **Introduction to data.** Data structures, variables, types of studies and experimental design and basic data collection techniques.
2. **Descriptive statistics.** Numerical measures, graphical representations of data, data summaries
3. **Probability.** The basic principles of probability.
4. **Distributions of random variables.** Commonly used distributions of discrete and continuous random variables. Introduction to the normal model and other key distributions.
5. **Basic Foundations for inference.** General ideas for statistical inference in the context of estimating the population mean. Emphasis on sampling theory.
6. **Confidence intervals.** One sample confidence intervals on the mean and on proportions; two sample confidence intervals on a difference of means on on a difference of proportions; confidence intervals on paired data.

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

Examples, exercises, and appendices

Examples and within-chapter exercises throughout the textbook may be identified by their distinctive bullets:

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

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

Exercise 0.2 Large empty bullets signal to readers that an exercise has been inserted into the text for additional practice and guidance. Students may find it useful to fill in the bullet after understanding or successfully completing the exercise. Solutions are provided for all within-chapter exercises in footnotes.[1]

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

OpenIntro, online resources, and getting involved

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

We encourage anyone learning or teaching statistics to visit **openintro.org** and get involved. We also provide many free online resources, including free course software. Data sets for this textbook are available on the website and through a companion R package.[2] All of these resources are free, and we want to be clear that anyone is welcome to use these online tools and resources with or without this textbook as a companion.

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

Acknowledgements

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

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

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

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

Chapter 0

Overview

Uncertainty is an inherent part of everyday life. We all face questions regarding uncertainty such as whether classed will go ahead as planned on any given day, will a flight leave on time, will a student pass a certain course? Uncertainties might also change depend on other factors, such as whether classes will still go ahead as planned when there is a snow warning in effect, if a flight is delayed can a person still manage to make their connection, will a student pass their course considering that the instructor is known to be a tough grader?

The ability to quantify uncertainty using rigorous mathematics is a powerful and useful tool. Calculating uncertainty on an intuitive level is something that is hard-wired in our DNA, such as the decision to fight or flight depending on the circumstances. However we cannot always make such intuitive decisions based purely on hunches and gut feelings. Fortunes have been lost based on someone having a good feeling on something. If we have some information available, we should make the best prediction possible using this information. For instance if we wanted to invest a lot of money in a company, we should use all available data such as past sales, market and industry trends, leadership ability of the CEO, forward looking statements etc. and with this information we can then predict whether our investment will be profitable.

In order for companies to survive and remain competitive in today's environment it is essential to monitor industry trends and read the market properly. Companies that don't adapt and stick to an outdated business model tend to pay the price. At the other end of the spectrum, companies that understand the needs of the consumer, build their product around the consumer and keep evolving their product offerings based on consumer trends tend to perform very well.

Statistics is the science of uncertainty and it is clearly a very useful subject for business. In this book you will be given an introduction to statistics and you will learn the framework as well as the language required at the introductory level. The material may be daunting at times, but the more you get familiar with the subject the more comfortable you become with it. As business students, doing well in a statistics course will give you a competitive edge since the ability to interpret and perform quantitative analytics are skills desired by many employers.

Chapter 1

Introduction to data

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

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

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

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

1.1 Case study: using stents to prevent strokes

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

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke.[1] Stents are devices put inside blood vessels that assist

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

in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

Does the use of stents reduce the risk of stroke?

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

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

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

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

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Table 1.1. Patient outcomes are recorded as "stroke" or "no event", 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

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

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

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

Table 1.2: Descriptive statistics for the stent study.

⊙ **Exercise 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 in-text exercises are provided using footnotes.)[2]

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

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

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

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

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

1.2 Data basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

Table 1.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 emails received during early 2012. These observations will be referred to as the `email50` data set, and they are a random sample from a larger data set that we will see in Section 2.3.

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

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

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

Table 1.3: Four rows from the email50 data matrix.

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

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

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

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

The data in Table 1.3 represent a **data matrix**, which is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A data matrix for the stroke study introduced in Section 1.1 is shown in Table 1.1 on page 3, where the cases were patients and there were three variables recorded for each patient.

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

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

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

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

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

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

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

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

variable	description
`name`	County name
`state`	State where the county resides (also including the District of Columbia)
`pop2000`	Population in 2000
`pop2010`	Population in 2010
`fed_spend`	Federal spending per capita
`poverty`	Percent of the population in poverty
`homeownership`	Percent of the population that lives in their own home or lives with the owner (e.g. children living with parents who own the home)
`multiunit`	Percent of living units that are in multi-unit structures (e.g. apartments)
`income`	Income per capita
`med_income`	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older
`smoking_ban`	Type of county-wide smoking ban in place at the end of 2011, which takes one of three values: `none`, `partial`, or `comprehensive`, where a `comprehensive` ban means smoking was not permitted in restaurants, bars, or workplaces, and `partial` means smoking was banned in at least one of those three locations

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

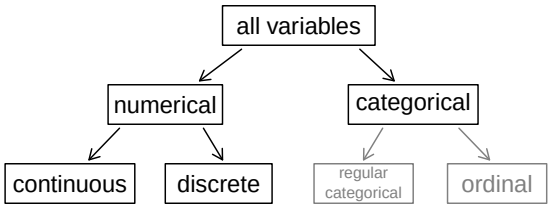


Figure 1.7: Breakdown of variables into their respective types.

1.2.2 Types of variables

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

First consider `fed_spend`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since their average, sum, and difference have no clear meaning.

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

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

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

Example 1.3 Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

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

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

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

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

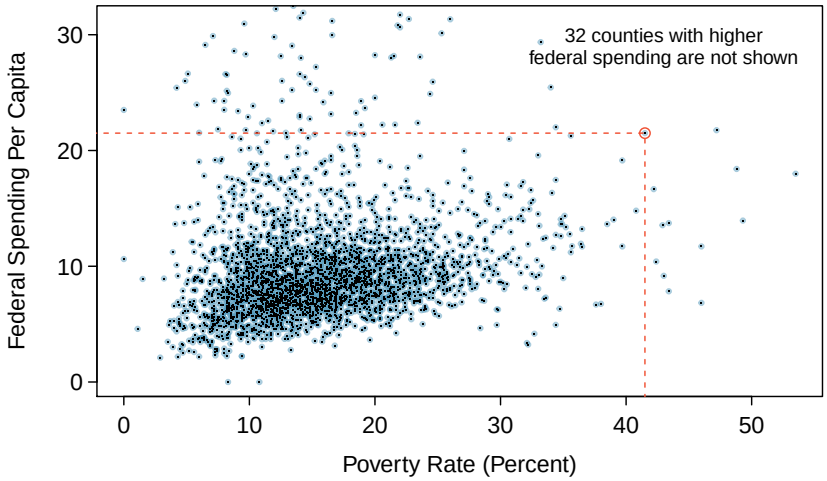


Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

1.2.3 Relationships between variables

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

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

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

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

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

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

⊙ **Exercise 1.5** Examine the variables in the `email50` data set, which are described in Table 1.4 on page 5. Create two questions about the relationships between these variables that are of interest to you.[9]

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

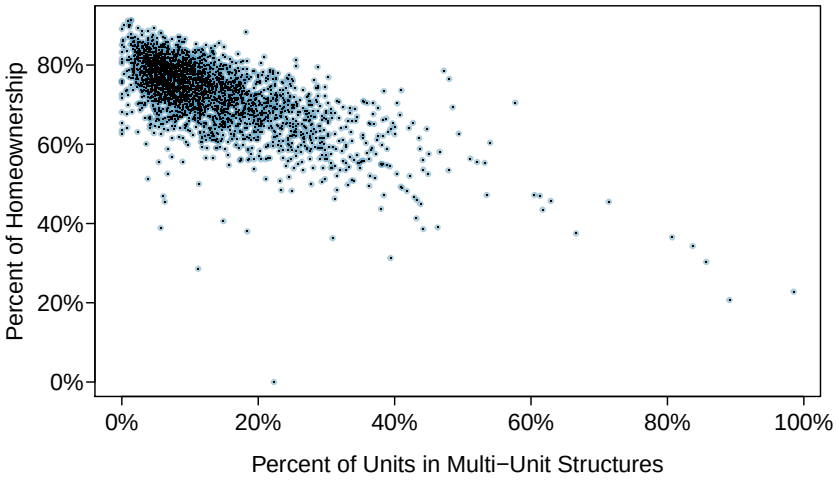


Figure 1.9: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties. Interested readers may find an image of this plot with an additional third variable, county population, presented at www.openintro.org/stat/down/MHP.png.

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

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

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

Because there is a downward trend in Figure 1.9 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `poverty` and `fed_spend` variables represented in Figure 1.8, where counties with higher poverty rates tend to receive more federal spending per capita.

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

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

1.3 Overview of data collection principles

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

1.3.1 Populations and samples

Consider the following three research questions:

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

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. 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. A single swordfish is a **unit**. A unit refers to objects that we are interested in studying. In statistical analysis measurements are recorded on units in a sample [10].

The key concept to grasp is that a population is large and it is usually very difficult to take measurements on every single unit in a population. A sample is much smaller than a population so we are able to take measurements on the units in a sample.

⊙ **Exercise 1.7** For the second and third questions above, identify the target population and what represents an individual case.[11]

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.

[10] Samples data may also contain incomplete measurements however techniques for working with missing data are more suited for a more advanced course.

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

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**.

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

February 10th, 2010.

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

1.3.3 Sampling from a population

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

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

● **Example 1.8** Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

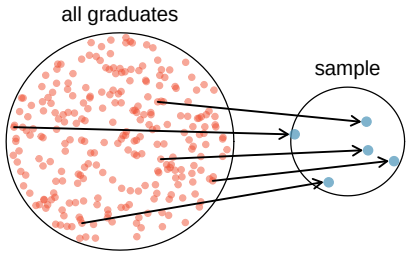


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

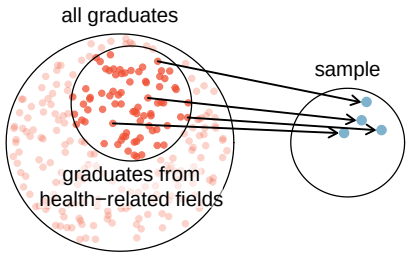


Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

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

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, and which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

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

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.

⊙ **Exercise 1.9** We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their

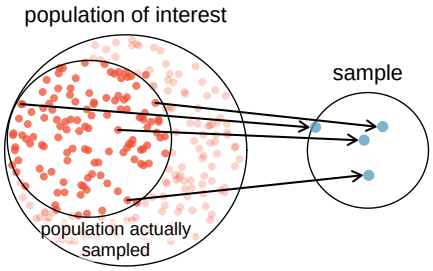


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[12]

1.3.4 Explanatory and response variables

Consider the following question from page 8 for the `county` data set:

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

If we suspect poverty might affect spending in a county, then poverty is the **explanatory** variable and federal spending is the **response** variable in the relationship.[13] If there are many variables, it may be possible to consider a number of them as explanatory variables.

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

explanatory variable → might affect → response variable

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

In some cases, there is no explanatory or response variable. Consider the following question from page 8:

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

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

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

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

1.3.5 Introducing observational studies and experiments

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

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to study why certain diseases might develop. In each of these situations, researchers merely observe 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.

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

1.4 Observational studies and sampling strategies

1.4.1 Observational studies

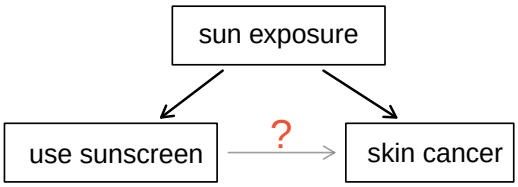
Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.

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

⊙ **Exercise 1.10** Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[14]

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

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.



Sun exposure is what is called a **confounding variable**,[15] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

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

⊙ **Exercise 1.11** Figure 1.9 shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest one or more other variables that might explain the relationship visible in Figure 1.9.[16]

Observational studies come in two forms: prospective and retrospective studies. A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of similar individuals over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The NursesÕ Health Study, started in 1976 and expanded in 1989.[17] This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets, such as `county`, may contain both prospectively- and retrospectively-collected variables. Local governments prospectively collect some variables as events unfolded (e.g. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

1.4.2 Three sampling methods (special topic)

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

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

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

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

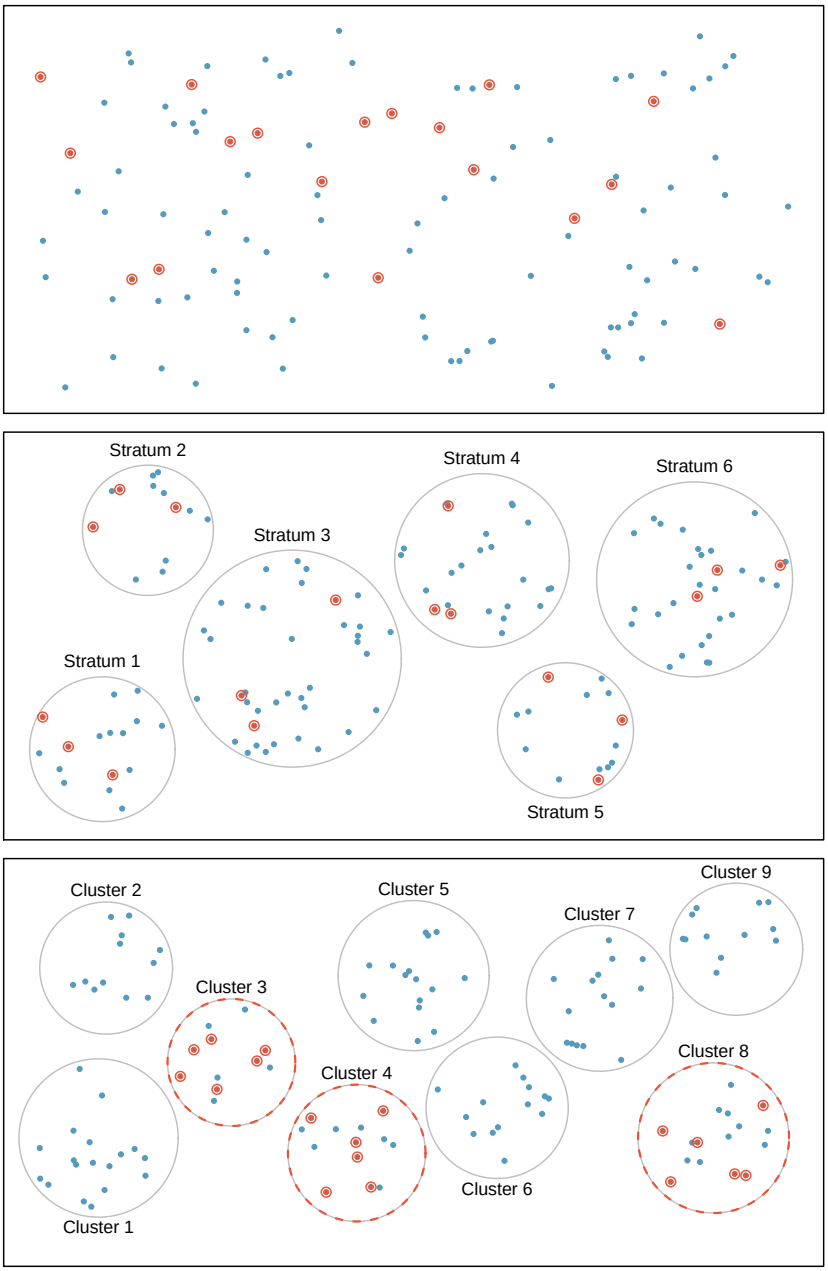


Figure 1.14: Examples of simple random, stratified, and cluster sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the middle panel, stratified sampling was used: cases were grouped into strata, and then simple random sampling was employed within each stratum.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

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries from the 2010 season, we could write the names of that season's 828 players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as "simple random" if each case in the population has an equal chance of being included in the final sample *and* knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata; some teams have a lot more money (we're looking at you, Yankees). Then we might randomly sample 4 players from each team for a total of 120 players.

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

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

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population.

A **cluster sample** is much like a two-stage simple random sample. We break up the population into many groups, called **clusters**. Then we sample a fixed number of clusters and collect a simple random sample within each cluster. This technique is similar to stratified sampling in its process, except that there is no requirement in cluster sampling to sample from every cluster. Stratified sampling requires observations be sampled from every stratum.

Sometimes cluster sampling can be a more economical random sampling technique than the alternatives. Also, unlike stratified sampling, cluster sampling is most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. For example, if neighborhoods represented clusters, then this sampling method works best when the neighborhoods are very diverse. A downside of cluster sampling is that more advanced analysis techniques are typically required, though the methods in this book can be extended to handle such data.

Example 1.13 Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However,

cluster sampling seems like a very good idea. First, we might randomly select half the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample and would still give us reliable information.

1.5 Experiments

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

1.5.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

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

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

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

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

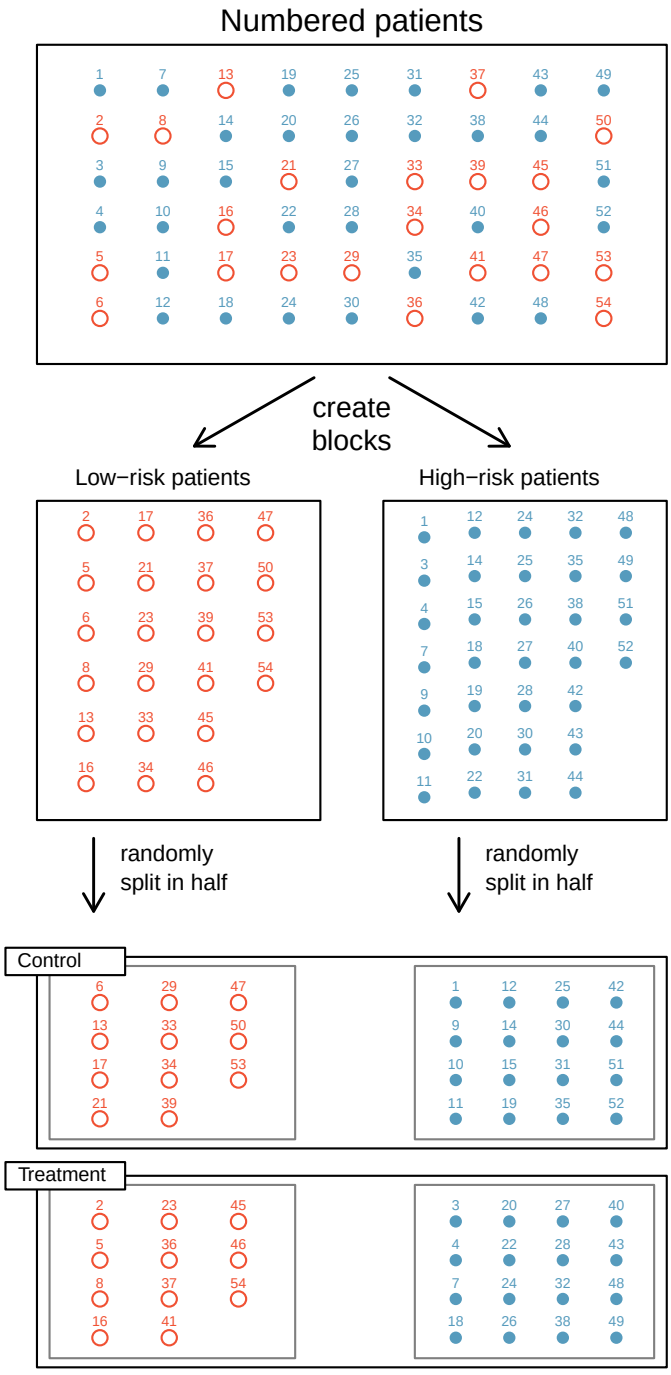


Figure 1.15: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both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

1.5.2 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients.[18] In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[19] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

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

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

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[20]

⊙ **Exercise 1.14** Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[21]

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

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

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

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

1.6 Introduction to statistical inference

One of the most important areas of statistics is statistical inference. With statistical inference we use statistical techniques to make generalizations on large groups based on information gathered from a small group. There are some preliminary definitions that are very important to get accustomed with.

We were introduced to populations and samples in Chapter 1.3.1. Along with populations and samples, two terms that we will be using a lot are "parameters" and "statistics". A **parameter** is numerical characteristic of a population. A **statistic** is a numerical characteristic of a sample. The goal of statistical inference is to use statistics to estimate parameters and to quantify the accuracy of these statistics with probabilities. In other words we want to quantify how sure we are about an estimate (i.e. a statistic) being equal to the truth (i.e. the parameter). Figure 1.16 illustrates populations and samples.

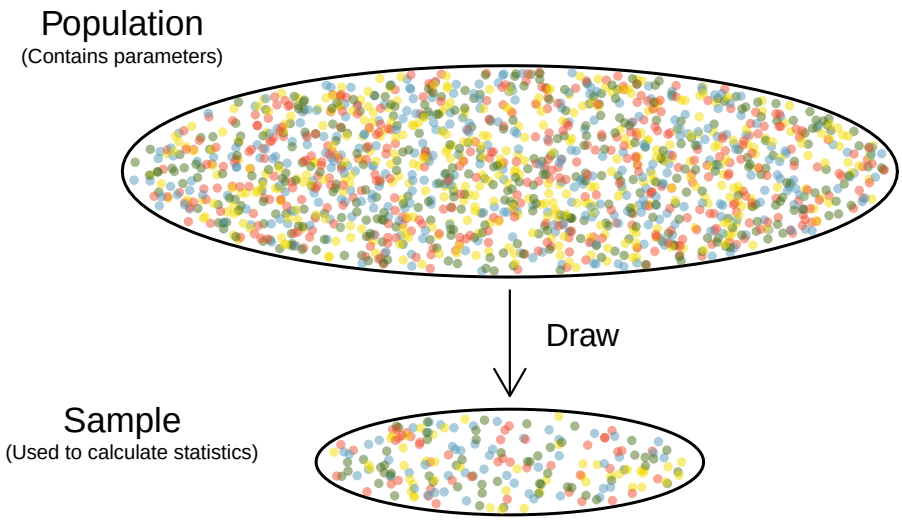


Figure 1.16: Figure illustrating populations and samples.

If we are studying a single population, the parameters of interest to us are the mean and standard deviation of the population or the population proportion. The mean is the average that we are used to working with and the standard deviation is a measure of spread relative to the mean. These terms are explained in more detail in Chapter 2.1.4 but we would like to introduce them at this point so that these terms become familiar to the reader. There are instances in which we may be interested in the proportion of units in a population rather than an average; such as the proportion of people who favour a particular political candidate, or the proportion of people who recover from a disease after taking a certain drug. A proportion is a fraction of the overall population.

Parameters are usually unknown since a population is very large. *Statistics are calculated and known* since a sample is much smaller than a population so we have a lot more control over the measurements in a sample.

In standard notation, parameters are associated with greek letters and statistics are

associated with lower case English letters. Table 1.17 below provides a summary of the standard notation used for parameters and statistics of a single population.

	Symbol	Description
Parameters	μ	Population mean
	σ	Population standard deviation
	p	Population proportion
Statistics	$\bar{x}$	Sample mean
	s	Sample standard deviation
	$\hat{p}$	Sample proportion

Table 1.17: Common notation for parameters and statistics of a single population

The symbol μ is pronounced *mew* and the symbol σ is pronounced *sigma*. Also $\bar{x}$ is called x *bar* and $\hat{p}$ is called p *hat*.

There is another parameter of interest which is the population variance σ^2 (and likewise we are also interested in the sample variance s^2), however the variance is the square of the standard deviation so including both variance and standard deviation in Table 1.17 is redundant. The variance and standard deviation are explained in more detail in Chapter 2.1.4.

If we were interested in two (or more) populations then we can use subscripts to denote the parameters and statistics of each individual population. For instance if we were interested in the difference between average male and average female height, we could label the population of males as 1 and the population of females as 2, so our parameter of interest would be $\mu_1 - \mu_2$. Likewise we can label the mean from a sample of males as $\bar{x}_1$ and the mean from a sample of females as $\bar{x}_2$ and calculate $\bar{x}_1$ - $\bar{x}_2$. We can apply similar labelling to the other parameters and statistics in Table 1.17.

There are many different sampling methods that can be applied to obtain sample data and the particular type of technique used may depend on factors the spatial layout or structure of the population, the amount of resources available to take measurements, the ease of making measurements and the amount of control we have in designing an experiment. In all cases we would ideally like to draw a **random sample** in which each unit in a population has an equal chance of being selected. Another characteristic in a random sample is that any combination n units also has an equal chance of being selected.

Quality of Sample Data
The quality of the estimates obtained depends heavily on the quality of the sample data

One fact that gets overlooked a lot is the quality of data that is used in statistical analysis. Statistical theory provides us with scientifically accepted tools to analyze data, however the quality of the results obtained depends on the quality of the data we have. Once we get accustomed to certain statistical methods, it may appear superficially as if the machinery is already in place, however if we input bad data we will end up with bad results. Therefore we stress the importance of obtaining good quality data for any statistical analysis.

Chapter 2

Descriptive Statistics

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

2.1 Numerical Measures

2.1.1 The Sample mean

The **mean**, sometimes called the average, is a common way to measure the (weighted) center of a **distribution** of data.

Mean

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

Let $x_1, x_2, \ldots, x_n$ represent the n observed values. The sample mean of these values is:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2.1}$$

n
sample size

We will calculate the mean for one variable in the `email50` data set. The variable (`num_char`) represents number of characters in emails for the `email50` data set. A pictorial way of representing number of characters this data is by using a dot plot. A dot plot is the most basic of displays. A **dot plot** is a one-variable scatterplot; an example using the number of characters from 50 emails is shown in Figure 2.1. A stacked version of this dot plot is shown in Figure 2.2. In a dot plot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figures 2.1 and 2.2. Note that some of the points

in 2.1 appear to overlap but this is due to the points being close together. We can always remove overlapping by scaling.

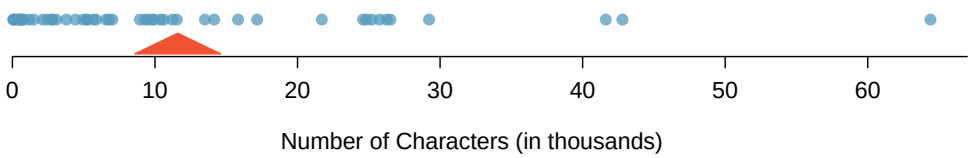


Figure 2.1: A dot plot of `num_char` for the `email50` data set.

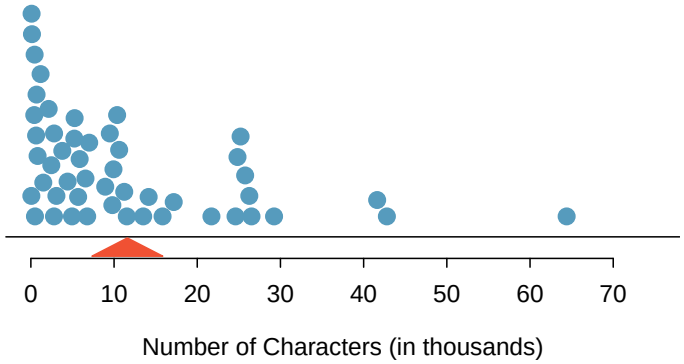


Figure 2.2: A stacked dot plot of `num_char` for the `email50` data set.

To find the mean number of characters in the 50 emails, we add up all the character counts and divide by the number of emails. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \tag{2.2}$$

$\bar{x}$ sample mean

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `num_char`, and the bar over on the x communicates that the average number of characters in the 50 emails was 11,600. It is useful to think of the mean as the balancing point of the distribution. The sample mean is shown as a triangle in Figures 2.1 and 2.2.

⊙ **Exercise 2.3** Examine Equations ((2.2)) and ((2.1)) above. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[1]

⊙ **Exercise 2.4** What was n in this sample of emails?[2]

The `email50` data set represents a sample from a larger population of emails that were received in January and March. We could compute a mean for this population in the same way as the sample mean, however, the population mean has a special label: μ. The symbol

μ population mean

[1] x_1 corresponds to the number of characters in the first email in the sample (21.7, in thousands), x_2 to the number of characters in the second email (7.0, in thousands), and x_i corresponds to the number of characters in the i^{th} email in the data set.

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

μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x.

Example 2.5 The average number of characters across all emails can be estimated using the sample data. Based on the sample of 50 emails, what would be a reasonable estimate of μ_x, the mean number of characters in all emails in the `email` data set? (Recall that `email50` is a sample from `email`.)

The sample mean, 11,600, may provide a reasonable estimate of μ_x. While this number will not be perfect, it provides a *point estimate* of the population mean. In Chapter 6 and beyond, we will develop tools to characterize the accuracy of point estimates, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

Example 2.6 We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,143 counties in the `county` data set. What would be a better approach?

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

Example (2.6) used what is called a **weighted mean**, which will not be a key topic in this textbook. However, we have provided an online supplement on weighted means for interested readers:

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

2.1.2 Median

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 n is *odd*

$$Median = \left(\frac{n+1}{2}\right) \text{ observation} \tag{2.7}$$

- If n is *even*

$$Median = \text{average of } \left(\frac{n}{2}\right) \text{ and } \left(\frac{n}{2}+1\right) \text{ observation} \tag{2.8}$$

If there are an even number of observations, there will be two values in the middle, and the median is taken as their average. When there are an odd number of observations,

there will be exactly one observation that splits the data into two halves, and in this case that observation is the median (no average needed).

2.1.3 Mode

Mode: most frequent values
Values in a data set that appear frequently relative to the rest of the data.

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

2.1.4 Variance and standard deviation

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

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

Let $x_1, x_2, \dots, x_n$ represent the n observed values. The sample variance of these values is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots (x_n - \bar{x})^2}{n-1} \tag{2.9}$$

The sample standard deviation of these values is:

$$s = \sqrt{s^2} \tag{2.10}$$

There are several reasons to explain the denominator of $n-1$ in Equation ((2.9)). In a more advanced statistics course, one of the most common explanations provided is that a denominator of $n-1$ makes s^2 and unbiased estimate of the population variance σ^2, however we will try to give a more intuitive explanation [3]. Notice that in the calculation of $\bar{x}$ in Equation ((2.1)) involved the use od n data points and then $\bar{x}$ is used in the calculation of s^2 so we are not really using all n values in calculating s^2 since s^2 depends on $\bar{x}$. This means we are losing a *degree of freedom* which is a concept that will be examined further in Section 7 and beyond.

[3] The formula for calculating the population variance is $s^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$. Here we divide by n instead of $n-1$ because we know the population mean

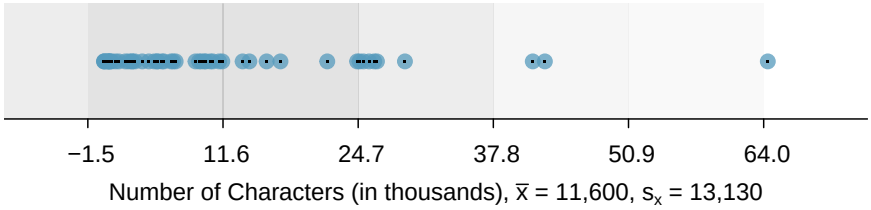


Figure 2.3: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

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

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

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

s^2
sample variance

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

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

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

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

s
sample standard deviation

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

TIP: standard deviation describes variability
Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 2.3 and 2.10, these percentages are not strict rules.

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 6 we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

2.2 Graphical Techniques

2.2.1 Histograms and shape

Dot plots show the exact value for each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. For example, in the `email50` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower bin. This tabulation is shown in Table 2.4. These binned counts are plotted as bars in Figure 2.7 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 2.2. A histogram is similar to a bar chart but for numerical data.

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

Table 2.4: The counts for the binned `num_char` data.

These bins that we divided our data into in Table 2.4 are also called **class intervals**. Class intervals can be equal or varying width. Once we count the occurrences at which the observed data falls into one our these class intervals and construct a **frequency table**. The general form of a frequency table is given in Table 2.5.

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
$[a_1, b_1)$	f_1	$r_1 = f_1/F$	f_1	r_1
$[a_2, b_2)$	f_2	$r_2 = f_2/F$	$f_1 + f_2$	$r_1 + r_2$
$[a_3, b_3)$	f_3	$r_3 = f_3/F$	$f_1 + f_2 + f_3$	$r_1 + r_2 + r_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_m, b_m]$	f_m	$r_m = f_m/F$	$f_1 + \ldots + f_m = F$	$r_1 + \ldots + r_m = 1$
	$F = \sum_{i=1}^{m} f_i$	1		

Table 2.5: General form of a frequency table.

To create a **frequency histogram**, the class intervals become the width of the bars

of the histogram and the frequencies becomes the heights. In a **Relative Frequency Histogram**, the class intervals become the width of the bars of the histogram and the relative frequencies become the heights. [4] The frequencies are the actual raw counts of data that were observed to fall inside a class interval. The relative frequencies are the proportion of data that was observed to fall inside a class interval.

The frequency table for the `num_char` data is given in Table 2.6.

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
$[0, 5)$	19	0.38	19	0.38
$[5, 10)$	12	0.24	31	0.62
$[10, 15)$	6	0.12	37	0.74
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[55, 60)$	0	0	49	0.98
$[60, 65)$	1	0.02	50	1
	50	1		

Table 2.6: Frequency table for `num_char` data.

The information in Table 2.6 was used to make the frequency histogram as well as the relative frequency histogram for `num_char` data. See Figures 2.7 and 2.8.

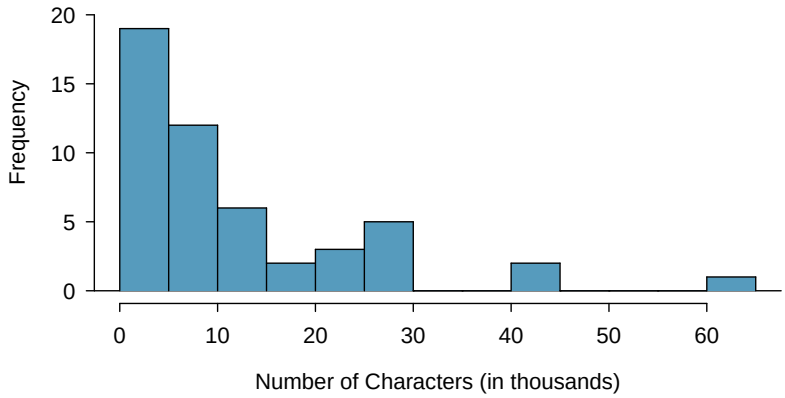


Figure 2.7: A frequency histogram of `num_char`. Notice that this distribution is very strongly skewed to the right.

[4] When we use the term "histogram" we are typically referring to a frequency histogram.

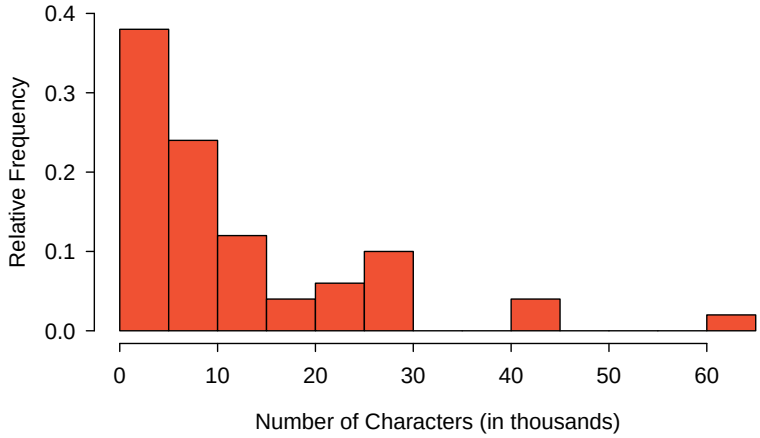


Figure 2.8: A relative frequency histogram of `num_char`. Notice how the shape is similar to the histogram in Figure 2.7.

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

Histograms are especially convenient for describing the shape of the data distribution. Figure 2.7 shows that most emails have a relatively small number of characters, while fewer emails have a very large number of characters. When data trail off to the right in this way and have a longer right tail, the shape is said to be **right skewed**.[5]

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

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

⊙ **Exercise 2.11** Take a look at the dot plots in Figures 2.1 and 2.2. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[6]

⊙ **Exercise 2.12** 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.

[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**.

[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]Character counts for individual emails.

A **mode** is represented by a prominent peak in the distribution. There is only one prominent peak in the histogram of `num_char`.

Figure 2.9 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.

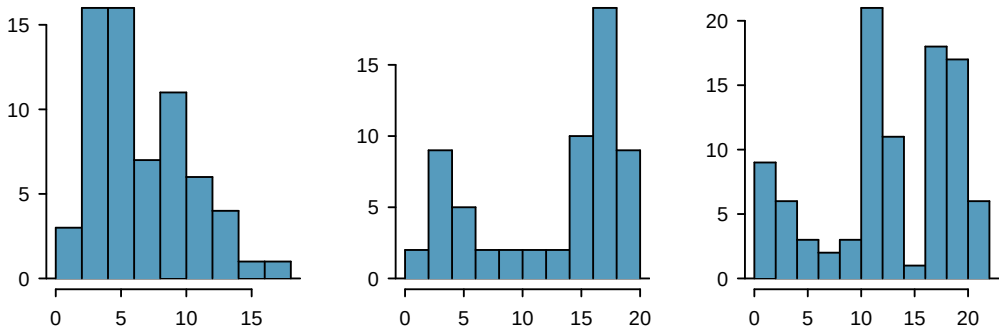


Figure 2.9: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

⊙ **Exercise 2.13** Figure 2.7 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[8]

⊙ **Exercise 2.14** Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?[9]

TIP: Looking for modes

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

⊙ **Exercise 2.15** On page 30, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.10 as an example, explain why such a description is important.[10]

[8] 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.)

[9] 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.

[10] Figure 2.10 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.

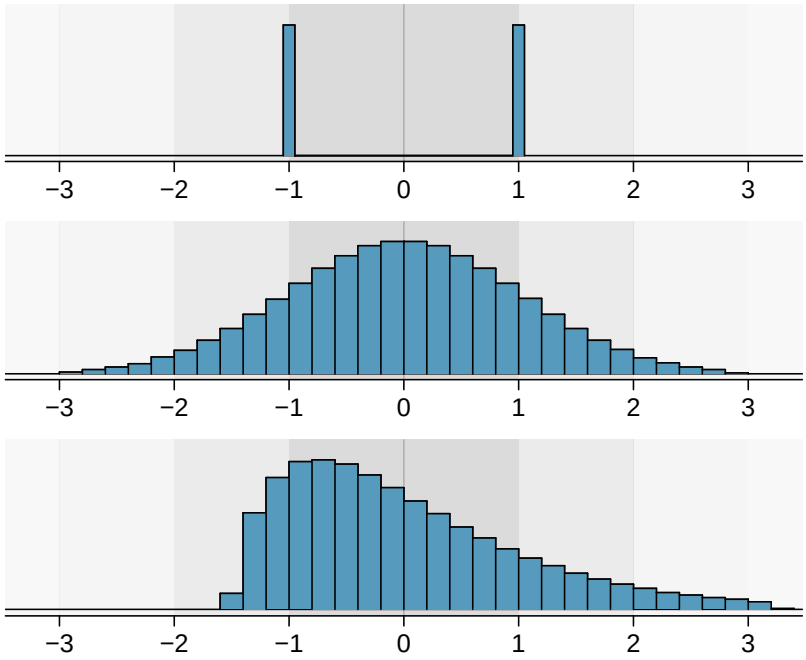


Figure 2.10: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

Example 2.16 Describe the distribution of the `num_char` variable using the histogram in Figure 2.7 on page 29. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.

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

2.2.2 Box plots and quartiles

A **box plot** summarizes a data set using five statistics while also plotting unusual observations. Figure 2.11 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email50` data set.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 2.11 shows 50% of the data falling below the median (dashes) and other 50% falling above the median (open circles). There are 50 character counts in the data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two observations closest to the 50^{th} percentile: $(6{,}768 + 7{,}012)/2 = 6{,}890$.

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 2.11, 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.

Figure 2.11: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The 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.

⊙ **Exercise 2.17** What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[11]

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 $1.5 \times IQR$.[12] They capture everything within this reach. In Figure 2.11, the upper whisker does not extend to the last three points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 33, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

[11] 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.
[12] While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.

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

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

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

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

⊙ **Exercise 2.18** The observation 64,401, a suspected outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[13]

⊙ **Exercise 2.19** Using Figure 2.11, estimate the following values for `num_char` in the `email50` data set: (a) Q_1, (b) Q_3, and (c) IQR.[14]

2.2.3 Robust statistics

How are the sample statistics of the `num_char` data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these summary statistics if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 2.12, and sample statistics are computed under each scenario in Table 2.13.

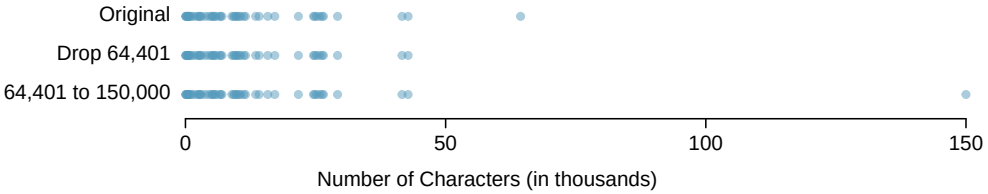


Figure 2.12: Dot plots of the original character count data and two modified data sets.

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

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

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

Table 2.13: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

⊙ **Exercise 2.20** (a) Which is more affected by extreme observations, the mean or median? Table 2.13 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[15]

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

● **Example 2.21** The median and IQR do not change much under the three scenarios in Table 2.13. Why might this be the case?

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

⊙ **Exercise 2.22** The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[16]

2.2.4 Transforming data (special topic)

When data are very strongly skewed, we sometimes transform them so they are easier to model. Consider the histogram of salaries for Major League Baseball players' salaries from 2010, which is shown in Figure 2.14(a).

● **Example 2.23** The histogram of MLB player salaries is useful in that we can see the data are extremely skewed and centered (as gauged by the median) at about $1 million. What isn't useful about this plot?

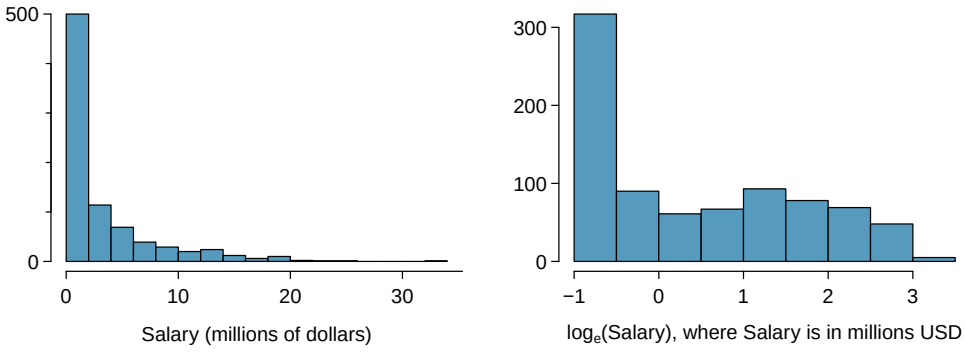
Most of the data are collected into one bin in the histogram and the data are so strongly skewed that many details in the data are obscured.

There are some standard transformations that are often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive. A **transformation** is a rescaling of the data using a function. For instance, a plot of the natural logarithm[17] of player salaries results in a new histogram in Figure 2.14(b).

[15] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Exercise (2.20).

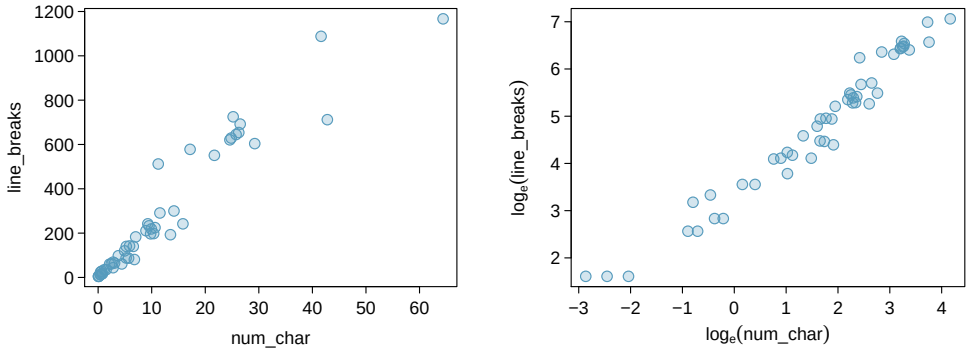
[16] Buyers of a "regular car" should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

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



(a) (b)

Figure 2.14: (a) Histogram of MLB player salaries for 2010, in millions of dollars. (b) Histogram of the log-transformed MLB player salaries for 2010.



(a) (b)

Figure 2.15: (a) Scatterplot of `line_breaks` against `num_char` for 50 emails. (b) A scatterplot of the same data but where each variable has been log-transformed.

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

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the `line_breaks` and `num_char` variables is shown in Figure 2.15(a), which was earlier shown in Figure **??**. We can see a positive association between the variables and that many observations are clustered near zero. In Chapter **??**, we might want to use a straight line to model the data. However, we'll find that the data in their current state cannot be modeled very well. Figure 2.15(b) shows a scatterplot where both the `line_breaks` and `num_char` variables have been transformed using a log (base e) transformation. While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.

Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are used by statisticians. Common goals in transforming data are to see the data structure differently, reduce skew, assist

in modeling, or straighten a nonlinear relationship in a scatterplot.

2.2.5 Mapping data (special topic)

The `county` data set offers many numerical variables that we could plot using dot plots, scatterplots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should map it using an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 2.16 and 2.17 shows intensity maps for federal spending per capita (`fed_spend`), poverty rate in percent (`poverty`), homeownership rate in percent (`homeownership`), and median household income (`med_income`). The color key indicates which colors correspond to which values. Note that the intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions.

Example 2.24 What interesting features are evident in the `fed_spend` and `poverty` intensity maps?

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

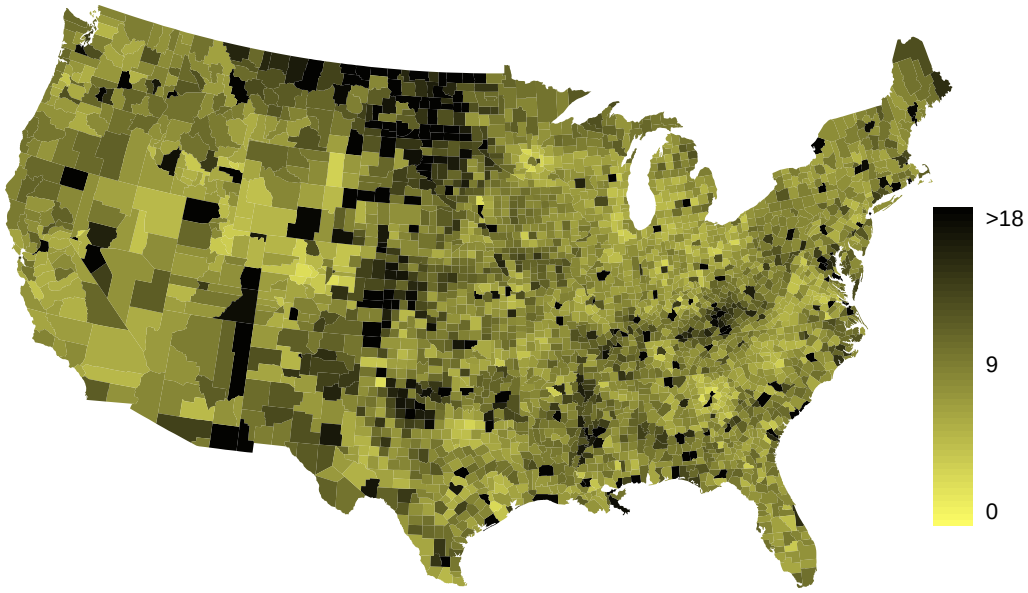
Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does the southwest border of Texas. The vertical strip of eastern Utah and Arizona, noted above for its higher federal spending, also appears to have higher rates of poverty (though generally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

Exercise 2.25 What interesting features are evident in the `med_income` intensity map?[18]

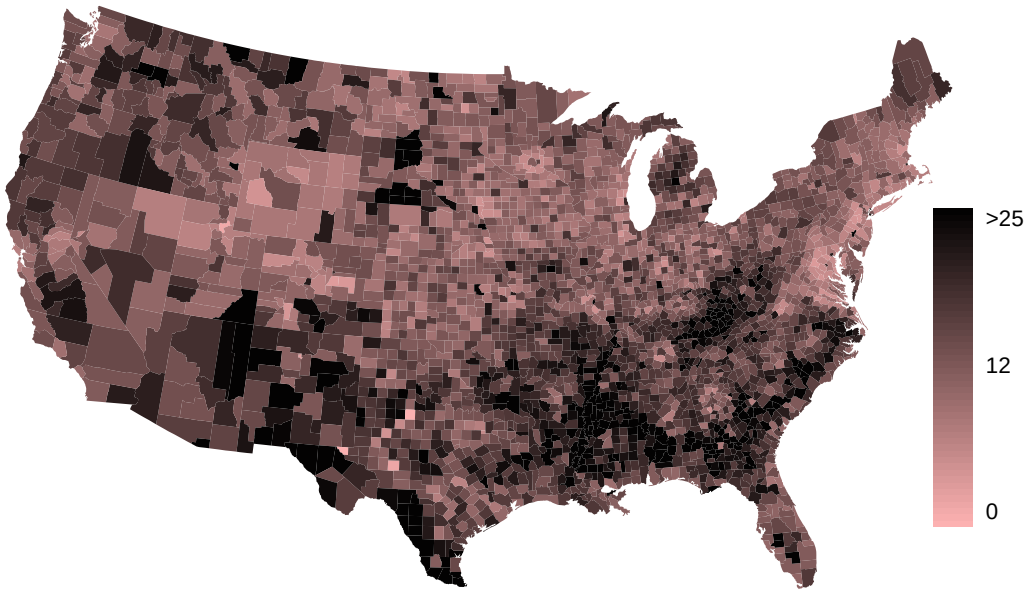
2.3 Considering categorical data

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

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

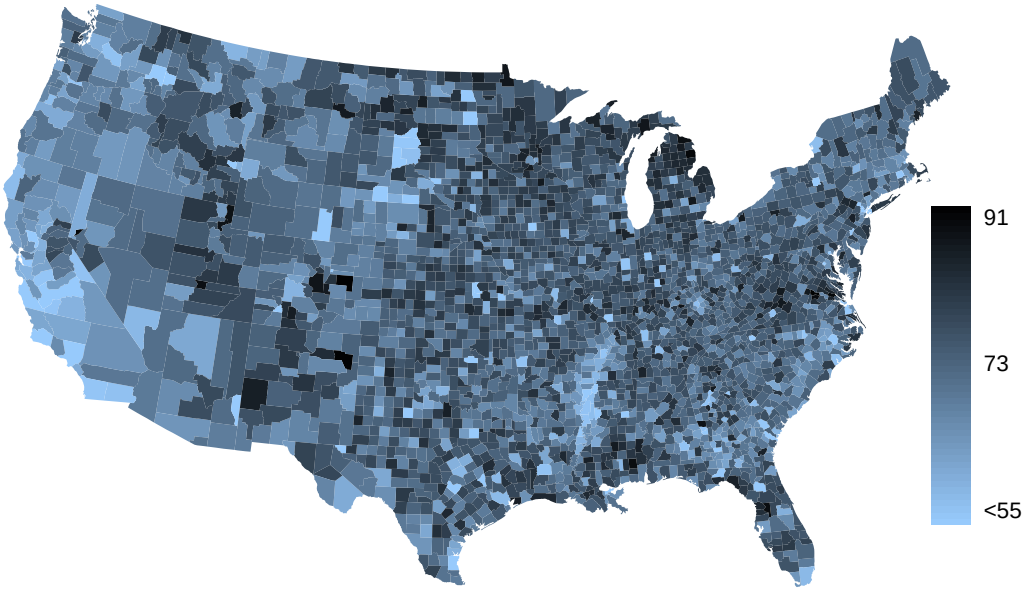


(a)

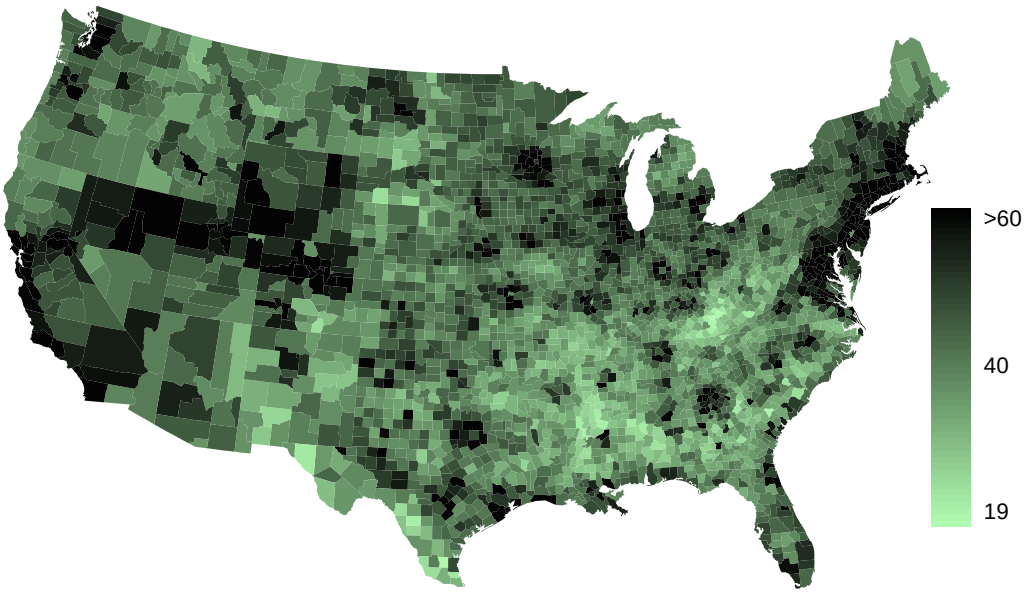


(b)

Figure 2.16: (a) Map of federal spending (dollars per capita). (b) Intensity map of poverty rate (percent).



(a)



(b)

Figure 2.17: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income ($1000s).

2.3.1 Contingency tables and bar plots

Table 2.18 summarizes two variables: `spam` and `number`. Recall that `number` is a categorical variable that describes whether an email contains no numbers, only small numbers (values under 1 million), or at least one big number (a value of 1 million or more). A table that summarizes data for two categorical variables in this way is called a **contingency table**. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 149 corresponds to the number of emails in the data set that are spam *and* had no number listed in the email. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $149 + 168 + 50 = 367$), and **column totals** are total counts down each column.

A table for a single variable is called a **frequency table**. Table 2.19 is a frequency table for the `number` variable. If we replaced the counts with percentages or proportions, the table would be called a **relative frequency table**.

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

Table 2.18: A contingency table for `spam` and `number`.

none	small	big	Total
549	2827	545	3921

Table 2.19: A frequency table for the `number` variable.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 2.20 shows a **bar plot** for the `number` variable. In the right panel, the counts are converted into proportions (e.g. $549/3921 = 0.140$ for `none`), showing the proportion of observations that are in each level (i.e. in each category).

2.3.2 Row and column proportions

Table 2.21 shows the row proportions for Table 2.18. The **row proportions** are computed as the counts divided by their row totals. The value 149 at the intersection of `spam` and `none` is replaced by $149/367 = 0.406$, i.e. 149 divided by its row total, 367. So what does 0.406 represent? It corresponds to the proportion of spam emails in the sample that do not have any numbers.

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

Table 2.21: A contingency table with row proportions for the `spam` and `number` variables.

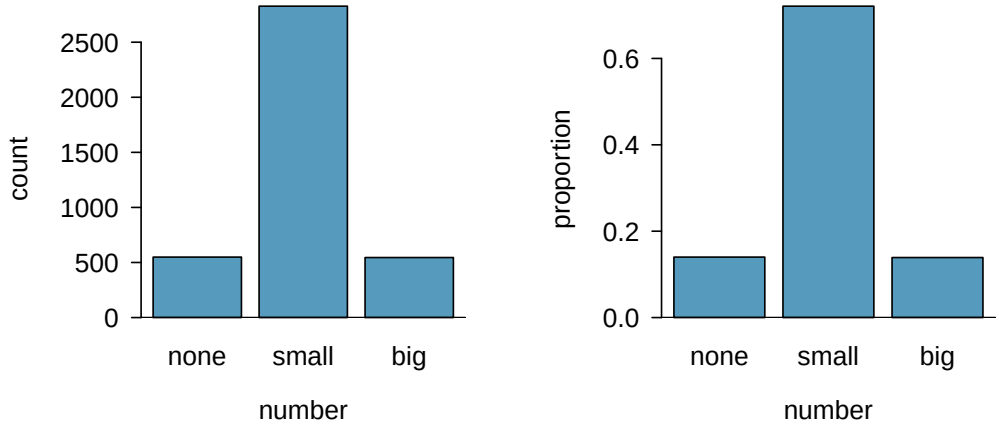


Figure 2.20: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Table 2.22 shows such a table, and here the value 0.271 indicates that 27.1% of emails with no numbers were spam. This rate of spam is much higher compared to emails with only small numbers (5.9%) or big numbers (9.2%). Because these spam rates vary between the three levels of `number` (`none`, `small`, `big`), this provides evidence that the `spam` and `number` variables are associated.

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

Table 2.22: A contingency table with column proportions for the `spam` and `number` variables.

We could also have checked for an association between `spam` and `number` in Table 2.21 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of emails with no numbers, small numbers, and big numbers varied from `spam` to `not spam`.

Exercise 2.26 What does 0.458 represent in Table 2.21? What does 0.059 represent in Table 2.22?[19]

Exercise 2.27 What does 0.139 at the intersection of `not spam` and `big` represent in Table 2.21? What does 0.908 represent in the Table 2.22?[20]

Example 2.28 Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content. A contingency table for the `spam` and `format` variables from the `email` data set are shown in Table 2.23. Recall that an HTML email is an email with the capacity for special formatting, e.g. bold text. In Table 2.23, which would be more helpful to someone hoping to classify email as spam or regular email: row or column proportions?

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

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

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

Table 2.23: A contingency table for `spam` and `format`.

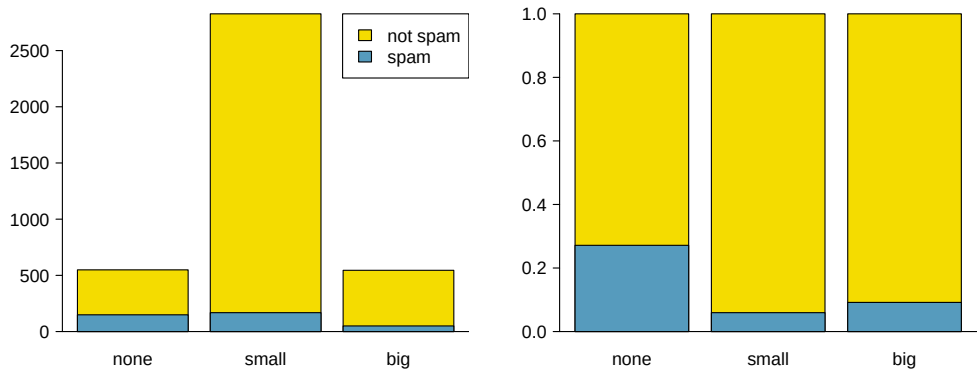
Example (2.28) points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed.

Exercise 2.29 Look back to Tables 2.21 and 2.22. Which would be more useful to someone hoping to identify spam emails using the `number` variable?[21]

[19] 0.458 represents the proportion of spam emails that had a small number. 0.058 represents the fraction of emails with small numbers that are spam.

[20] 0.139 represents the fraction of non-spam email that had a big number. 0.908 represents the fraction of emails with big numbers that are non-spam emails.

[21] The column proportions in Table 2.22 will probably be most useful, which makes it easier to see that emails with small numbers are spam about 5.9% of the time (relatively rare). We would also see that about 27.1% of emails with no numbers are spam, and 9.2% of emails with big numbers are spam.



(a) (b)

Figure 2.24: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by `spam`. (b) Standardized version of Figure (a).

2.3.3 Segmented bar and mosaic plots

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

A **segmented bar plot** is a graphical display of contingency table information. For example, a segmented bar plot representing Table 2.22 is shown in Figure 2.24(a), where we have first created a bar plot using the `number` variable and then divided each group by the levels of `spam`. The column proportions of Table 2.22 have been translated into a standardized segmented bar plot in Figure 2.24(b), which is a helpful visualization of the fraction of spam emails in each level of `number`.

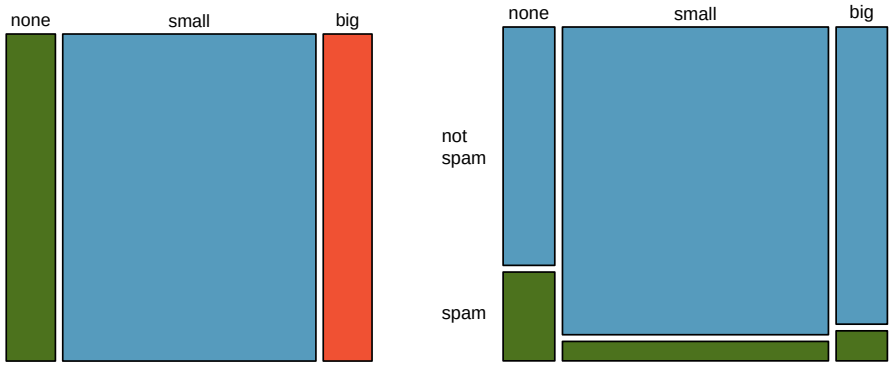
Example 2.30 Examine both of the segmented bar plots. Which is more useful?

Figure 2.24(a) contains more information, but Figure 2.24(b) presents the information more clearly. This second plot makes it clear that emails with no number have a relatively high rate of spam email – about 27%! On the other hand, less than 10% of email with small or big numbers are spam.

Since the proportion of spam changes across the groups in Figure 2.24(b), we can conclude the variables are dependent, which is something we were also able to discern using table proportions. Because both the `none` and `big` groups have relatively few observations compared to the `small` group, the association is more difficult to see in Figure 2.24(a).

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

A **mosaic plot** is a graphical display of contingency table information that is similar to a bar plot for one variable or a segmented bar plot when using two variables. Figure 2.25(a) shows a mosaic plot for the `number` variable. Each column represents a level of `number`, and the column widths correspond to the proportion of emails of each number type. For instance, there are fewer emails with no numbers than emails with only small numbers, so



(a) (b)

Figure 2.25: The one-variable mosaic plot for `number` and the two-variable mosaic plot for both `number` and `spam`.

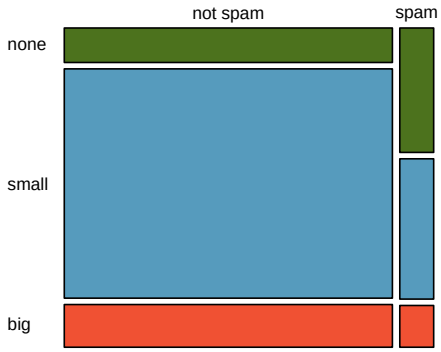


Figure 2.26: Mosaic plot where emails are grouped by the `number` variable after they've been divided into `spam` and `not spam`.

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

This one-variable mosaic plot is further divided into pieces in Figure 2.25(b) using the `spam` variable. Each column is split proportionally according to the fraction of emails that were spam in each number category. For example, the second column, representing emails with only small numbers, was divided into emails that were spam (lower) and not spam (upper). As another example, the bottom of the third column represents spam emails that had big numbers, and the upper part of the third column represents regular emails that had big numbers. We can again use this plot to see that the `spam` and `number` variables are associated since some columns are divided in different vertical locations than others, which was the same technique used for checking an association in the standardized version of the segmented bar plot.

In a similar way, a mosaic plot representing row proportions of Table 2.18 could be constructed, as shown in Figure 2.26. However, because it is more insightful for this application to consider the fraction of spam in each category of the `number` variable, we prefer Figure 2.25(b).

2.3.4 Pie charts

While pie charts are well known, they are not typically as useful as other charts in a data analysis. A **pie chart** is shown in Figure 2.27 alongside a bar plot. It is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions. In the case of the `none` and `big` categories, the difference is so slight you may be unable to distinguish any difference in group sizes for either plot!

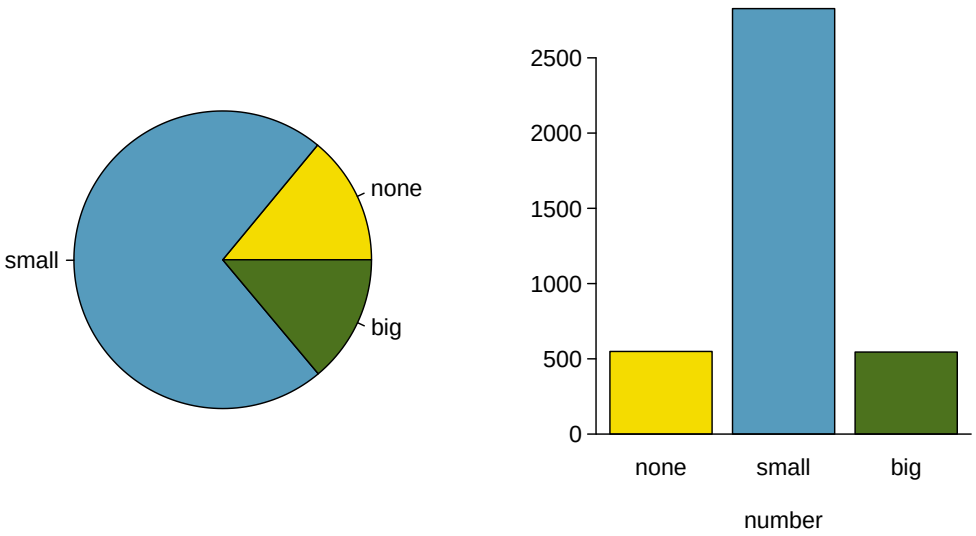


Figure 2.27: A pie chart and bar plot of `number` for the `email` data set.

2.3.5 Comparing numerical data across groups

Some of the more interesting investigations can be considered by examining numerical data across groups. The methods required here aren't really new. All that is required is to make a numerical plot for each group. Here two convenient methods are introduced: side-by-side box plots and hollow histograms.

We will take a look again at the `county` data set and compare the median household income for counties that gained population from 2000 to 2010 versus counties that had no gain. While we might like to make a causal connection here, remember that these are observational data and so such an interpretation would be unjustified.

There were 2,041 counties where the population increased from 2000 to 2010, and there were 1,099 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Table 2.28 to give a better sense of some of the raw data.

The **side-by-side box plot** is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 2.29, 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 2.29.

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

Table 2.28: In this table, median household income (in $1000s) from a random sample of 100 counties that gained population over 2000-2010 are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

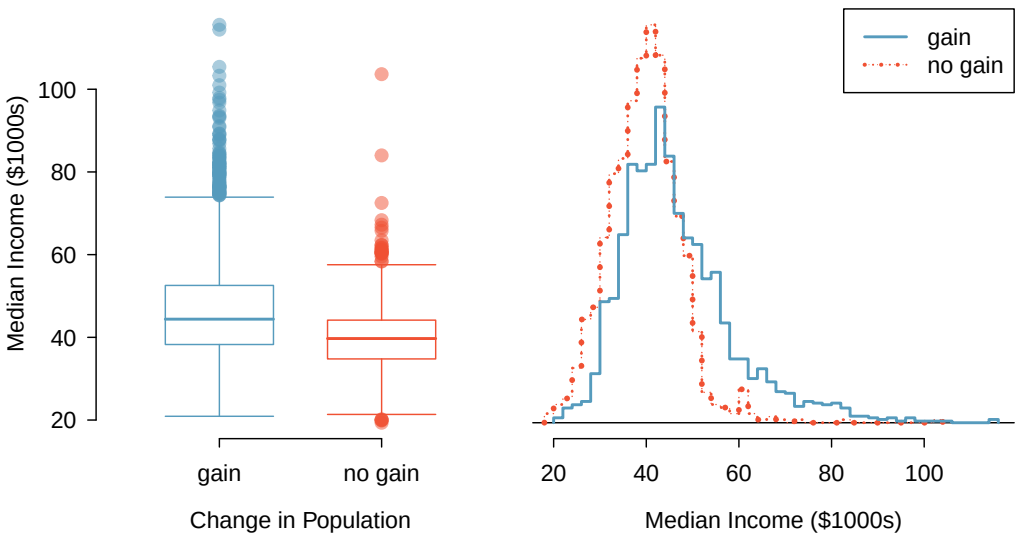


Figure 2.29: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.

⊙ **Exercise 2.31** Use the plots in Figure 2.29 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?[22]

⊙ **Exercise 2.32** What components of each plot in Figure 2.29 do you find most useful?[23]

2.4 Case study: gender discrimination (special topic)

● **Example 2.33** Suppose your professor splits the students in class into two groups: students on the left and students on the right. If $\hat{p}_L$ and $\hat{p}_R$ represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if $\hat{p}_L$ did not exactly equal $\hat{p}_R$?

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

⊙ **Exercise 2.34** 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?[24]

2.4.1 Variability within data

We consider a study investigating gender discrimination in the 1970s, which is set in the context of personnel decisions within a bank.[25] The research question we hope to answer is, "Are females unfairly discriminated against in promotion decisions made by male managers?"

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

[22] Answers may vary a little. The counties with population gains tend to have higher income (median of about $45,000) versus counties without a gain (median of about $40,000). The variability is also slightly larger for the population gain group. This is evident in the IQR, which is about 50% bigger in the *gain* group. Both distributions show slight to moderate right skew and are unimodal. There is a secondary small bump at about $60,000 for the *no gain* group, visible in the hollow histogram plot, that seems out of place. (Looking into the data set, we would find that 8 of these 15 counties are in Alaska and Texas.) The box plots indicate there are many observations far above the median in each group, though we should anticipate that many observations will fall beyond the whiskers when using such a large data set.

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

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

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

Exercise 2.35 Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[26]

For each supervisor we record the gender associated with the assigned file and the promotion decision. Using the results of the study summarized in Table 2.30, we would like to evaluate if females are unfairly discriminated against in promotion decisions. In this study, a smaller proportion of females are promoted than males (0.583 versus 0.875), but it is unclear whether the difference provides *convincing evidence* that females are unfairly discriminated against.

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

Table 2.30: Summary results for the gender discrimination study.

Example 2.36 Statisticians are sometimes called upon to evaluate the strength of evidence. When looking at the rates of promotion for males and females in this study, what comes to mind as we try to determine whether the data show convincing evidence of a real difference?

The observed promotion rates (58.3% for females versus 87.5% for males) suggest there might be discrimination against women in promotion decisions. However, we cannot be sure if the observed difference represents discrimination or is just from random chance. Generally there is a little bit of fluctuation in sample data, and we wouldn't expect the sample proportions to be *exactly* equal, even if the truth was that the promotion decisions were independent of gender.

Example (2.36) is a reminder that the observed outcomes in the sample may not perfectly reflect the true relationships between variables in the underlying population. Table 2.30 shows there were 7 fewer promotions in the female group than in the male group, a difference in promotion rates of 29.2% $\left(\frac{21}{24} - \frac{14}{24} = 0.292\right)$. This difference is large, but the sample size for the study is small, making it unclear if this observed difference represents discrimination or whether it is simply due to chance. We label these two competing claims, H_0 and H_A:

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

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

What would it mean if the independence model, which says the variables `gender` and `decision` are unrelated, is true? It would mean each banker was going to decide whether

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

to promote the candidate without regard to the gender indicated on the file. That is, the difference in the promotion percentages was due to the way the files were randomly divided to the bankers, and the randomization just happened to give rise to a relatively large difference of 29.2%.

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

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

2.4.2 Simulating the study

Table 2.30 shows that 35 bank supervisors recommended promotion and 13 did not. Now, suppose the bankers' decisions were independent of gender. Then, if we conducted the experiment again with a different random arrangement of files, differences in promotion rates would be based only on random fluctuation. We can actually perform this **randomization**, which simulates what would have happened if the bankers' decisions had been independent of gender but we had distributed the files differently.

In this **simulation**, we thoroughly shuffle 48 personnel files, 24 labeled `male_sim` and 24 labeled `female_sim`, and deal these files into two stacks. We will deal 35 files into the first stack, which will represent the 35 supervisors who recommended promotion. The second stack will have 13 files, and it will represent the 13 supervisors who recommended against promotion. Then, as we did with the original data, we tabulate the results and determine the fraction of `male_sim` and `female_sim` who were promoted. The randomization of files in this simulation is independent of the promotion decisions, which means any difference in the two fractions is entirely due to chance. Table 2.31 show the results of such a simulation.

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

Table 2.31: Simulation results, where any difference in promotion rates between `male_sim` and `female_sim` is purely due to chance.

⊙ **Exercise 2.37** What is the difference in promotion rates between the two simulated groups in Table 2.31? How does this compare to the observed 29.2% in the actual groups?[27]

[27] $18/24 - 17/24 = 0.042$ or about 4.2% in favor of the men. This difference due to chance is much smaller than the difference observed in the actual groups.

2.4.3 Checking for independence

We computed one possible difference under the independence model in Exercise (2.37), which represents one difference due to chance. While in this first simulation, we physically dealt out files, it is more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance: -0.042. And another: 0.208. And so on until we repeat the simulation enough times that we have a good idea of what represents the *distribution of differences from chance alone*. Figure 2.32 shows a plot of the differences found from 100 simulations, where each dot represents a simulated difference between the proportions of male and female files that were recommended for promotion.

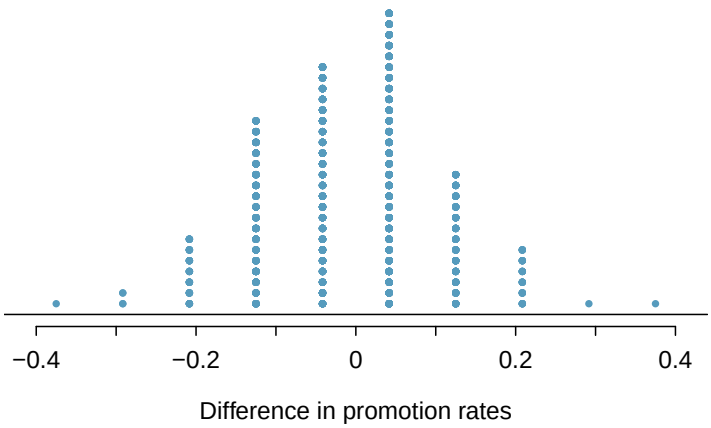


Figure 2.32: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where `gender_sim` and `decision` are independent. Two of the 100 simulations had a difference of at least 29.2%, the difference observed in the study.

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

Example 2.38 How often would you observe a difference of at least 29.2% (0.292) according to Figure 2.32? Often, sometimes, rarely, or never?

It appears that a difference of at least 29.2% due to chance alone would only happen about 2% of the time according to Figure 2.32. Such a low probability indicates a rare event.

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

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

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

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong evidence against the independence model. That is, we do not have sufficiently strong evidence to conclude there was gender discrimination. (2) We conclude the evidence is sufficiently strong to reject H_0 and assert that there was gender discrimination. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event.[28] So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence of gender discrimination against women by the supervisors.

One field of statistics, statistical inference, is built on evaluating whether such differences are due to chance. In statistical inference, statisticians evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur. In Chapter 6, 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.

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

Chapter 3

Probability

Probability forms a foundation for statistics. You might already be familiar with many aspects of probability, however, formalization of the concepts is new for most. In a probability problem certain characteristics of the population, such as the values of parameters are either known or at least assumed to be known. We then use this information to answer questions regarding observing specific outcome associated with a sample drawn from that population. This chapter aims to introduce probability on familiar terms using processes most people have seen before.

3.1 Defining probability

3.1.1 Probability

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

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

We start by defining an "experiment" in the context of probability. This term with regards to probability is not quite the same as with every day English.

Experiment
An experiment is an activity in which the outcome is not known for certain until the activity is completed.

In probability theory, rolling a die or flipping a coin are examples of experiments.

Sample Space
A sample space is the set of all possible outcomes of an experiment.

Sample Point
A sample point is an element of of a sample space.

A sample space describes any and all outcomes that can occur. A sample space can be a finite set or an infinite set. The symbol Ω is often used to denote a sample space. The experiment of rolling a die produces a value in the set {1, 2, 3, 4, 5, 6}. This set of all possible outcomes is the sample space (Ω) for rolling a die. The outcomes 1, 2, 3, 4, 5, 6 are sample points.

S
Sample space

Event
An event is any collection of sample points.

In the example of rolling a die, we can consider an event to be obtaining a 3 on a roll as an event or obtaining an even number on a roll as an event. With this framework in place, we can attach probabilities to events occurring within a sample space. We now give the formal definition of probability.

Probability
The **probability** of an event is the proportion of times the event 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%.

Example 3.1 A "die", the singular of dice, is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting 1 when rolling a die?

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

Example 3.2 What is the chance of getting a 1 or 2 in the next roll?

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

Example 3.3 What is the chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?

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

Example 3.4 What is the chance of not rolling a 2?

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

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

Example 3.5 Consider rolling two dice. If $1/6^{th}$ of the time the first die is a 1 and $1/6^{th}$ of those times the second die is a 1, what is the chance of getting two 1s?

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

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**.

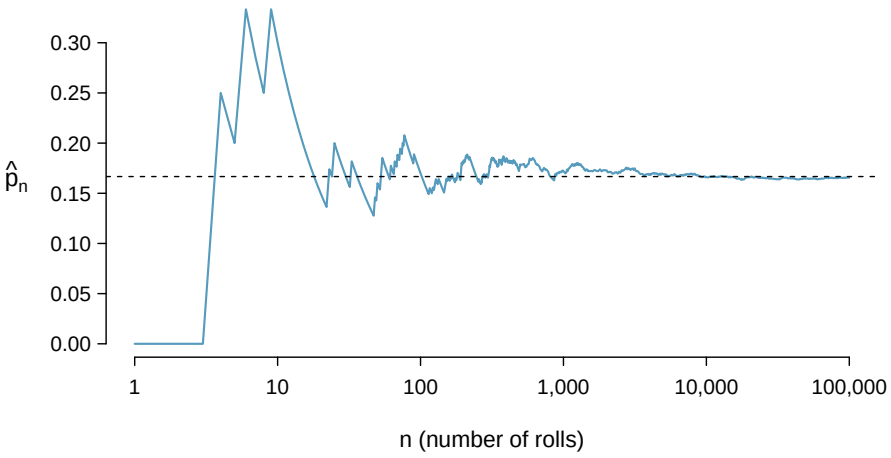


Figure 3.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

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.

Example 3.6 Consider p as the probability of rolling a `1` (as above). We can also write this probability as

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

$P(A)$
Probability of outcome A

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 $P(\texttt{1})$.

Exercise 3.7 Random processes include rolling a die and flipping a coin. (a) Think of another random process. (b) Describe all the possible outcomes of that process.

For instance, rolling a die is a random process with potential outcomes `1`, `2`, ..., `6`.[1]

What we think of as random processes are not necessarily random, but they may just be too difficult to understand exactly. The fourth example in the footnote solution to Exercise (3.7) suggests a roommate's behavior is a random process. However, even if a roommate's behavior is not truly random, modeling her behavior as a random process can still be useful.

TIP: Modeling a process as random
It can be helpful to model a process as random even if it is not truly random.

3.1.2 Common Notation

We always write statements regarding probabilities with capital P along with something meaningful within the context of a probability problem in parentheses such as in in Example (3.6). The information in parentheses could be an event or a sample point or a combination of events etc. We can write long statements in words but it is more elegant to use certain symbols. A list of common symbols is given in Table 3.2.

Common symbols in set theory

$\cap$	:	Intersection
$\cup$	:	Union
$\square^c$	:	Complement
$\subset$	:	Completely contained
$\subseteq$	:	Contained
$\in$	:	Is an element of
$\notin$	:	Is not an element of

Table 3.2: Common notation used in set theory.

The symbols in Table 3.2 become very useful when we work with events and write statements for probability problems.

Their interaction refers to elements that are in both A and B. Their union refers to elements that are in A or B. The complement of set A refers to all elements except for those in set A. We talk more about complements in Section 3.1.6. If $A \subset B$, this means that B contains all the elements of A as well as some additional elements. If $A \subseteq B$, this means that A is equal to B or B contains all the elements of A as well as some additional elements. If we write $x \in A$ this means that x is an element of A and if we write $x \notin A$ means that x is not an element of A.

[1] Here are four examples. (i) Whether someone gets sick in the next month or not is an apparently random process with outcomes `sick` and `not`. (ii) We can *generate* a random process by randomly picking a person and measuring that person's height. The outcome of this process will be a positive number. (iii) Whether the stock market goes up or down next week is a seemingly random process with possible outcomes `up`, `down`, and `no_change`. Alternatively, we could have used the percent change in the stock market as a numerical outcome. (iv) Whether your roommate cleans her dishes tonight probably seems like a random process with possible outcomes `cleans_dishes` and `leaves_dishes`.

Common symbols in set theory with events and sample points
Let A and B be any 2 events in sample space Ω.

$A \cap B$	:	Common to both A and B
$A \cup B$	:	Everything in A as well as B
A^c	:	Everything except A
$A \subset B$	:	A is completely contained in B
$A \subseteq B$	:	A is contained in B
$x \in A$	:	x is an element of A
$x \notin A$	:	x is not an element of A

Table 3.3: Common notation used in set theory when we have two events.

3.1.3 Venn Diagrams

Venn diagrams are a pictorial way to represent a sample space, events and sample points. They are useful when outcomes can be categorized as "in" or "out" for two or three variables, attributes, or random processes. Venn diagrams are an effective way of easily illustrating the concepts of set notation in Table 3.2 or 3.3 such as intersections, unions and complements.

Let A and B be any 2 events in sample space Ω. A few examples of representing A and B along with some of the symbols in Table 3.2 are given in Figures 3.4, 3.5 and 3.6 below.

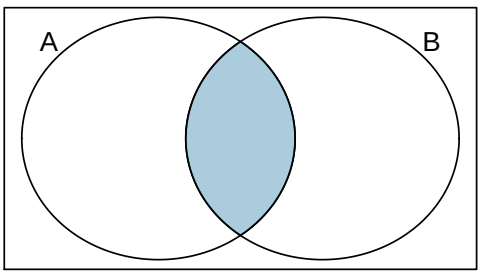


Figure 3.4: Shaded area represents $A \cap B$

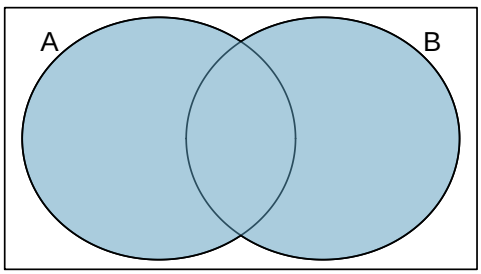


Figure 3.5: Shaded area represents $A \cup B$

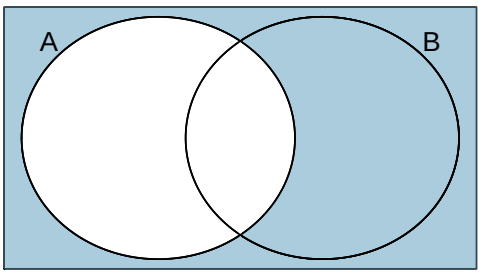


Figure 3.6: Shaded area represents A^c

We can use all sorts of combinations of the symbols in Table 3.2 to represent useful statements regarding probabilities.

3.1.4 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 and B represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A \text{ or } B) = P(A) + P(B) \tag{3.8}$$

In set notation

$$P(A \cup B) = P(A) + P(B) \tag{3.9}$$

⊙ **Exercise 3.10** 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]

⊙ **Exercise 3.11** In the `email` data set in Chapter 2, the `number` variable described whether no number (labeled `none`), only one or more small numbers (`small`), or whether at least one big number appeared in an email (`big`). Of the 3,921 emails, 549 had no numbers, 2,827 had only one or more small numbers, and 545 had at least one big number. (a) Are the outcomes `none`, `small`, and `big` disjoint? (b) Determine the proportion of emails with value `small` and `big` separately. (c) Use the Addition

[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) $P(\texttt{1 or 4 or 5}) = P(\texttt{1}) + P(\texttt{4}) + P(\texttt{5}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

Rule for disjoint outcomes to compute the probability a randomly selected email from the data set has a number in it, small or big.[3]

Statisticians rarely work with individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where a die roll results in `1` or `2` and B represent the event that the die roll is a `4` or a `6`. We write A as the set of outcomes $\{1, 2\}$ and $B = \{4, 6\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 3.7.

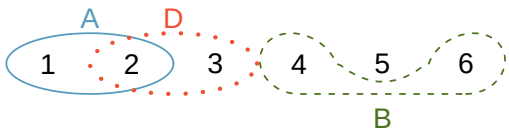


Figure 3.7: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

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:

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

Exercise 3.12 (a) Verify the probability of event A, $P(A)$, is $1/3$ using the Addition Rule. (b) Do the same for event B.[4]

Exercise 3.13 (a) Using Figure 3.7 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]

Exercise 3.14 In Exercise (3.13), you confirmed B and D from Figure 3.7 are disjoint. Compute the probability that either event B or event D occurs.[6]

3.1.5 Probabilities when events are not disjoint

Let's consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Table 3.8. If you are unfamiliar with the cards in a regular deck, please see the footnote.[7]

Exercise 3.15 (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 email is categorized in only one level of `number`. (b) Small: $\frac{2827}{3921} = 0.721$. Big: $\frac{545}{3921} = 0.139$. (c) $P(\texttt{small or big}) = P(\texttt{small}) + P(\texttt{big}) = 0.721 + 0.139 = 0.860$.

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

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

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

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

[8] (a) There are 52 cards and 13 diamonds. If the cards are thoroughly shuffled, each card has an equal

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

Table 3.8: Representations of the 52 unique cards in a deck.

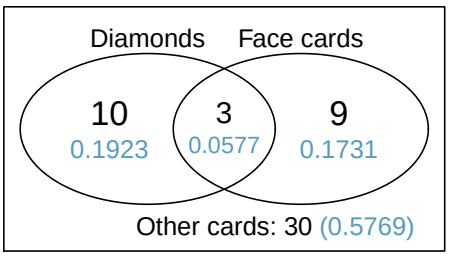


Figure 3.9: Venn diagram for diamonds and face cards.

The Venn diagram in Figure 3.9 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. $10/52 = 0.1923$).

⊙ **Exercise 3.16** Using the Venn diagram, verify $P(\text{face card}) = 12/52 = 3/13$.[9]

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

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

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

$$\begin{aligned} P(A \text{ or } B) &= P(\text{face card or } \diamondsuit) \\ &= P(\text{face card}) + P(\diamondsuit) - P(\text{face card and } \diamondsuit) \\ &= 12/52 + 13/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned} \tag{3.17}$$

Equation ((3.17)) is an example of the **General Addition Rule**.

chance of being drawn, so the probability that a randomly selected card is a diamond is $P(\diamondsuit) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

[9]The Venn diagram shows face cards split up into "face card but not $\diamondsuit$" and "face card and $\diamondsuit$". Since these correspond to disjoint events, $P(\text{face card})$ is found by adding the two corresponding probabilities: $\frac{3}{52} + \frac{9}{52} = \frac{12}{52} = \frac{3}{13}$.

General Addition Rule

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

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{3.18}$$

In set notation

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.19}$$

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.

⊙ **Exercise 3.20** (a) If A and B are disjoint, describe why this implies $P(A \text{ and } B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the simpler Addition Rule for disjoint events if A and B are disjoint.[10]

⊙ **Exercise 3.21** In the `email` data set with 3,921 emails, 367 were spam, 2,827 contained some small numbers but no big numbers, and 168 had both characteristics. Create a Venn diagram for this setup.[11]

⊙ **Exercise 3.22** (a) Use your Venn diagram from Exercise (3.21) to determine the probability a randomly drawn email from the `email` data set is spam and had small numbers (but not big numbers). (b) What is the probability that the email had either of these attributes?[12]

3.1.6 Complement of an event

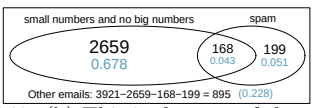
We often use the sample space to examine the scenario where an event does not occur. Recall that the complement of an event is the case in which the event does not occur. For example, let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D represents 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 3.10 shows the relationship between D, D^c, and the sample space Ω.

A^c
Complement of outcome A

⊙ **Exercise 3.23** (a) Compute $P(D^c) = P(\text{rolling a } 1, 4, 5, \text{ or } 6)$. (b) What is $P(D) + P(D^c)$?[13]

[10] (a) If A and B are disjoint, A and B can never occur simultaneously. (b) If A and B are disjoint, then the last term of Equation ((3.18)) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

[11] Both the counts and corresponding probabilities (e.g. 2659/3921 = 0.678) are shown. Notice that the number of emails represented in the left circle corresponds to $2659 + 168 = 2827$, and the number represented in the right circle is $168 + 199 = 367$.



[12] (a) The solution is represented by the intersection of the two circles: 0.043. (b) This is the sum of the three disjoint probabilities shown in the circles: $0.678 + 0.043 + 0.051 = 0.772$.

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

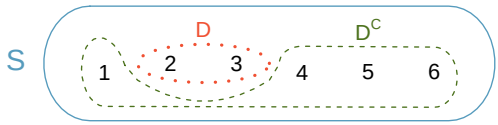


Figure 3.10: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. Ω represents the sample space, which is the set of all possible events.

⊙ **Exercise 3.24** Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 3.7 on page 58. (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:

1. Every possible outcome not in A is in A^c
2. A and A^c are disjoint

Property 1 implies

$$P(A \text{ or } A^c) = 1 \tag{3.25}$$

i.e.

$$P(A \cup A^c) = 1 \tag{3.26}$$

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

$$P(A \text{ or } A^c) = P(A) + P(A^c) \tag{3.27}$$

i.e.

$$P(A \cup A^c) = P(A) + P(A^c) \tag{3.28}$$

Combining Equations ((3.25)) and ((3.27)) 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:

$$P(A) + P(A^c) = 1 \tag{3.29}$$

An equivalent but more useful way to write expression (3.29) is

$$P(A^c) = 1 - P(A) \tag{3.30}$$

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.

[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 $P(B) + P(B^c) = 1$.

⊙ **Exercise 3.31** Let A represent the event where we roll two dice and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Table 3.11 on page 62. (c) Determine $P(A)$.[15]

⊙ **Exercise 3.32** Consider again the probabilities from Table 3.11 and rolling two dice. Find the following probabilities: (a) The sum of the dice is *not* 6. (b) The sum is at least 4. That is, determine the probability of the event $B = \{4, 5, ..., 12\}$. (c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, ..., 10\}$.[16]

3.1.7 Probability distributions

A **probability distribution** is a table (i.e. a function) of all disjoint outcomes and their associated probabilities.

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.

Table 3.11 shows the probability distribution for the sum of two dice.

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

Table 3.11: Probability distribution for the sum of two dice.

There are some special probability distributions that we will learn about in later sections of this text in Section 3.4 on random variables.

⊙ **Exercise 3.33** Table 3.12 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?[17]

Chapter 2 emphasized the importance of plotting data to provide quick summaries. Probability distributions can also be summarized in a bar plot. For instance, the distri-

[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 Equation ((3.29)): $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

[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 $P(D) = 1 - P(D^c) = 11/12$.

[17] 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).

Income range ($1000s)	0-25	25-50	50-100	100+
(a)	0.18	0.39	0.33	0.16
(b)	0.38	-0.27	0.52	0.37
(c)	0.28	0.27	0.29	0.16

Table 3.12: Proposed distributions of US household incomes (Exercise (3.33)).

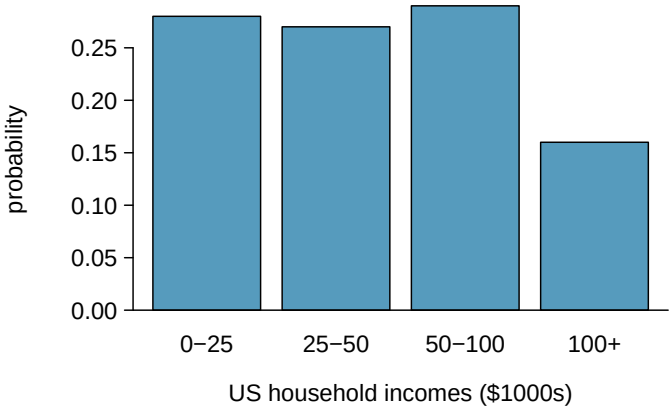


Figure 3.13: The probability distribution of US household income.

bution of US household incomes is shown in Figure 3.13 as a bar plot.[18] The probability distribution for the sum of two dice is shown in Table 3.11 and plotted in Figure 3.14.

In these bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a bar plot that resembles a histogram, as in the case of the sum of two dice. Another example of plotting the bars at their respective locations is shown in Figure 3.27 on page 88.

3.2 Conditional probability and Independence

Are students more likely to use marijuana when their parents used drugs? The `drug_use` data set contains a sample of 445 cases with two variables, `student` and `parents`, and is summarized in Table 3.15.[19] The `student` variable is either `uses` or `not`, where a student is labeled as `uses` if she has recently used marijuana. The `parents` variable takes the value `used` if at least one of the parents used drugs, including alcohol.

Example 3.34 If at least one parent used drugs, what is the chance their child (`student`) uses?

We will estimate this probability using the data. Of the 210 cases in this data set

[18] It is also possible to construct a distribution plot when income is not artificially binned into four groups. *Continuous* distributions are considered in Section 3.4.2.

[19] Ellis GJ and Stone LH. 1979. Marijuana Use in College: An Evaluation of a Modeling Explanation. Youth and Society 10:323-334.

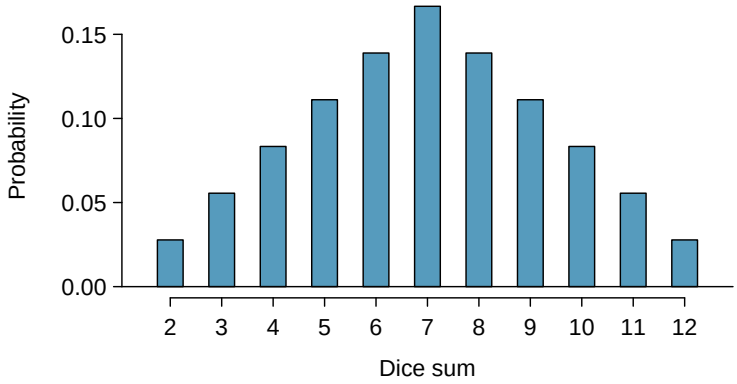


Figure 3.14: The probability distribution of the sum of two dice.

		parents		
		used	not	Total
student	uses	125	94	219
	not	85	141	226
	Total	210	235	445

Table 3.15: Contingency table summarizing the drug_use data set.

where parents = used, 125 represent cases where student = uses:

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60$$

Example 3.35 A student is randomly selected from the study and she does not use drugs. What is the probability that at least one of her parents used?

If the student does not use drugs, then she is one of the 226 students in the second row. Of these 226 students, 85 had at least one parent who used drugs:

$$P(\texttt{parents} = \texttt{used} \text{ given } \texttt{student} = \texttt{not}) = \frac{85}{226} = 0.376$$

3.2.1 Marginal and joint probabilities

Table 3.17 includes row and column totals for each variable separately in the drug_use data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without conditioning on any other variables. For instance, a probability based solely on the student variable is a marginal probability:

$$P(\texttt{student} = \texttt{uses}) = \frac{219}{445} = 0.492$$

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

$$P(\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{not}) = \frac{94}{445} = 0.21$$

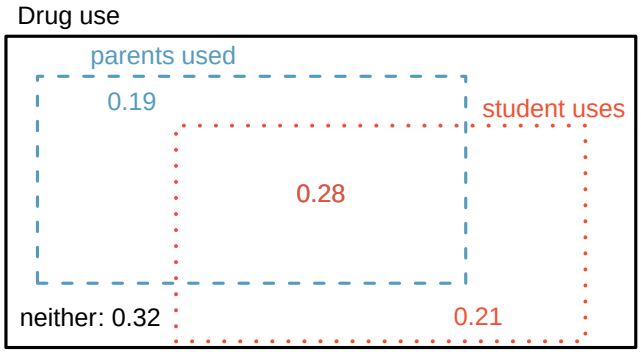


Figure 3.16: A Venn diagram using boxes for the `drug_use` data set.

	`parents: used`	`parents: not`	Total
`student: uses`	0.28	0.21	0.49
`student: not`	0.19	0.32	0.51
Total	0.47	0.53	1.00

Table 3.17: Probability table summarizing parental and student drug use.

It is common to substitute a comma for "and" in a joint probability, although either is acceptable.

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

We use **table proportions** to summarize joint probabilities for the `drug_use` sample. These proportions are computed by dividing each count in Table 3.15 by 445 to obtain the proportions in Table 3.17. The joint probability distribution of the `parents` and `student` variables is shown in Table 3.18.

⊙ **Exercise 3.36** Verify Table 3.18 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[20]

[20] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.28 + 0.19 + 0.21 + 0.32 = 1.00$.

Joint outcome	Probability
`parents` = `used`, `student` = `uses`	0.28
`parents` = `used`, `student` = `not`	0.19
`parents` = `not`, `student` = `uses`	0.21
`parents` = `not`, `student` = `not`	0.32
Total	1.00

Table 3.18: A joint probability distribution for the `drug_use` data set.

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability a random student from the study uses drugs is found by summing the outcomes from Table 3.18 where `student` = `uses`:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses}) \\ &\quad = P(\texttt{parents} = \texttt{used}, \texttt{student} = \texttt{uses}) + \\ &\qquad\quad P(\texttt{parents} = \texttt{not}, \texttt{student} = \texttt{uses}) \\ &\quad = 0.28 + 0.21 = 0.49 \end{aligned}$$

3.2.2 Defining conditional probability

Probabilities can change when a situation changes. If we have more information regarding a scenario then the probability of an event may change. For example the probability of a student taking a course might increase if they find out that their instructor has a reputation for giving easy tests. As such we have the notion of conditional probability.

Conditional Probability

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

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{3.37}$$

In set notation

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.38}$$

$P(A|B)$
Probability of outcome A given B

The expression $P(A|B)$ is the probability of event A occurring given (i.e on the condition) that event B has occurred. The vertical bar "|" is read as *given*. We are essentially examining a restricting the sample space in which our focus shifts from all of Ω to just B. The only way that A can occur in our restricted sample space of B is if we examine $P(A \cap B)$. We will discuss this concept some more with an example.

There is some connection between drug use of parents and of the student: drug use of one is associated with drug use of the other.[21] In this section, we discuss how to use information about associations between two variables to improve probability estimation.

The probability that a random student from the study uses drugs is 0.49. Could we update this probability if we knew that this student's parents used drugs? Absolutely. To do so, we limit our view to only those 210 cases where parents used drugs and look at the fraction where the student uses drugs:

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60$$

This is a **conditional probability** because we computed the probability under a condition: `parents` = `used`. There are two parts to a conditional probability, **the outcome of interest** and the **condition**. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event.

We separate the text inside our probability notation into the outcome of interest and the condition:

[21] This is an observational study and no causal conclusions may be reached.

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used})$$
$$= P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60 \tag{3.39}$$

In Equation ((3.39)), we computed the probability a student uses based on the condition that at least one parent used as a fraction:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) \\ &\quad = \frac{\#\text{ times } \texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used}}{\#\text{ times } \texttt{parents} = \texttt{used}} \\ &\quad = \frac{125}{210} = 0.60 \end{aligned} \tag{3.40}$$

We considered only those cases that met the condition, $\texttt{parents} = \texttt{used}$, and then we computed the ratio of those cases that satisfied our outcome of interest, the student uses.

Counts are not always available for data, and instead only marginal and joint probabilities may be provided. For example, disease rates are commonly listed in percentages rather than in a count format. We would like to be able to compute conditional probabilities even when no counts are available, and we use Equation ((3.40)) as an example demonstrating this technique.

We considered only those cases that satisfied the condition, $\texttt{parents} = \texttt{used}$. Of these cases, the conditional probability was the fraction who represented the outcome of interest, $\texttt{student} = \texttt{uses}$. Suppose we were provided only the information in Table 3.17 on page 65, i.e. only probability data. Then if we took a sample of 1000 people, we would anticipate about 47% or $0.47 \times 1000 = 470$ would meet our information criterion. Similarly, we would expect about 28% or $0.28 \times 1000 = 280$ to meet both the information criterion and represent our outcome of interest. Thus, the conditional probability could be computed:

$$\begin{aligned} P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) &= \frac{\#\ (\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used})}{\#\ (\texttt{parents} = \texttt{used})} \\ &= \frac{280}{470} = \frac{0.28}{0.47} = 0.60 \end{aligned} \tag{3.41}$$

In Equation ((3.41)), we examine exactly the fraction of two probabilities, 0.28 and 0.47, which we can write as

$$P(\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used}) \quad \text{and} \quad P(\texttt{parents} = \texttt{used}).$$

The fraction of these probabilities represents our general formula for conditional probability.

⊙ **Exercise 3.42** (a) Write out the following statement in conditional probability notation: "*The probability a random case has* $\texttt{parents} = \texttt{not}$ *if it is known that* $\texttt{student} = \texttt{not}$". Notice that the condition is now based on the student, not the parent. (b) Determine the probability from part (a). Table 3.17 on page 65 may be helpful.[22]

⊙ **Exercise 3.43** (a) Determine the probability that one of the parents had used drugs if it is known the student does not use drugs. (b) Using the answers from part (a) and Exercise (3.42)(b), compute

$$P(\texttt{parents} = \texttt{used} | \texttt{student} = \texttt{not}) + P(\texttt{parents} = \texttt{not} | \texttt{student} = \texttt{not})$$

[22] (a) $P(\texttt{parent} = \texttt{not} | \texttt{student} = \texttt{not})$. (b) Equation ((3.37)) for conditional probability indicates we should first find $P(\texttt{parents} = \texttt{not} \text{ and } \texttt{student} = \texttt{not}) = 0.32$ and $P(\texttt{student} = \texttt{not}) = 0.51$. Then the ratio represents the conditional probability: $0.32/0.51 = 0.63$.

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

Table 3.19: Contingency table for the smallpox data set.

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

Table 3.20: Table proportions for the smallpox data, computed by dividing each count by the table total, 6224.

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

Exercise 3.44 The data indicate that drug use of parents and children are associated. Does this mean the drug use of parents causes the drug use of the students?[24]

3.2.3 Smallpox in Boston, 1721

The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[25] 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 3.19 and 3.20.

Exercise 3.45 Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[26]

[23] (a) This probability is $\frac{P(\texttt{parents = used and student = not})}{P(\texttt{student = not})} = \frac{0.19}{0.51} = 0.37$. (b) The total equals 1. (c) Under the condition the student does not use drugs, the parents must either use drugs or not. The complement still appears to work *when conditioning on the same information.*

[24] No. This was an observational study. Two potential confounding variables include income and region. Can you think of others?

[25] Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6)*. Geneva: World Health Organization. ISBN 92-4-156110-6.

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

⊙ **Exercise 3.46** Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Exercise (3.45)?[27]

⊙ **Exercise 3.47** 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?[28]

3.2.4 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.

Statistical independence
Let A and B be any 2 events in sample space Ω. Events A and B are independent if and only if

$$P(A|B) = P(A) \quad (3.48)$$
$$P(B|A) = P(B) \quad (3.49)$$
$$P(A \cap B) = P(A) \cdot P(B) \quad (3.50)$$

If 2 events are independent then (3.48), (3.49) and (3.50) must all occur. 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 53), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a `1`, and $1/6^{th}$ of *those* times the white die will also be `1`. This is illustrated in Figure 3.21. 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.

The main idea is that if two processes are independent, then knowing the outcome of one should provide no information about the other and we can show this is mathematically true using conditional probabilities.

⊙ **Exercise 3.51** Let X and Y represent the outcomes of rolling two dice. (a) What is the probability that the first die, X, is `1`? (b) What is the probability that both X and Y are `1`? (c) Use the formula for conditional probability to compute $P(Y =$

[27] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{yes}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{yes})}{P(\texttt{inoculated} = \texttt{yes})} = \frac{0.0010}{0.0392} = 0.0255$. The death rate for individuals who were inoculated is only about 1 in 40 while the death rate is about 1 in 7 for those who were not inoculated.

[28] 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).

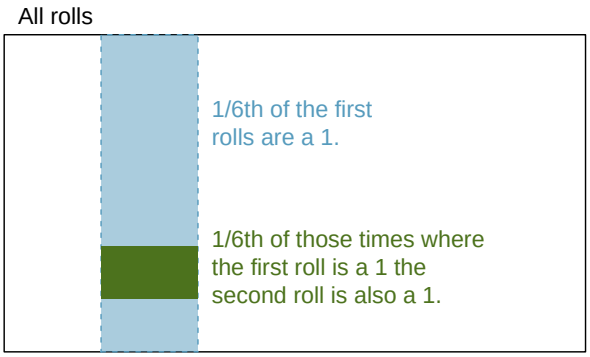


Figure 3.21: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

1 $|X = 1$). (d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.[29]

We can show in Exercise (3.51)(c) that the conditioning information has no influence by using the Multiplication Rule for independence processes:

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

⊙ **Exercise 3.52** Ron is watching a roulette table in a casino and notices that the last five outcomes were black. He figures that the chances of getting black six times in a row is very small (about 1/64) and puts his paycheck on red. What is wrong with his reasoning?[30]

● **Example 3.53** What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all 1s?

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

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

[29] 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.

[30] 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**.

Examples (3.5) and (3.53) illustrate what is called the Multiplication Rule for independent processes.

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

$$P(A \text{ and } B) = P(A) \cdot P(B) \tag{3.54}$$

In set notation

$$P(A \cup B) = P(A) \cdot P(B) \tag{3.55}$$

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

$$P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_k)$$

⊙ **Exercise 3.56** 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?[31]

⊙ **Exercise 3.57** Suppose 5 people are selected at random.[32]

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

Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[33] using

[31] (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 $0.91 \times 0.91 = 0.8281$.

[32] (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:

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

(b) Using the same reasoning as in (a), $0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$
(c) Use the complement, $P(\text{all five are } \texttt{RH})$, to answer this question:

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

[33] 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.

the Multiplication Rule:

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

Exercise 3.58 Three people are selected at random.[34]

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

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

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

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

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation ((3.54)).

Example 3.59 If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is 1/4 and the probability that it is an ace is 1/13. The probability the card is the ace of hearts is 1/52. We check whether Equation (3.54) 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.

3.2.5 General multiplication rule

Section 3.2.4 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

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

In set notation

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

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

This General Multiplication Rule is simply a rearrangement of the definition for conditional probability in Equation ((3.37)) on page 66.

[34] 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.

Example 3.60 Consider the `smallpox` data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

We will compute our answer using the General Multiplication Rule and then verify it using Table 3.20. We want to determine

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

and we are given that

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

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

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

This is equivalent to the General Multiplication Rule. We can confirm this probability in Table 3.20 at the intersection of `no` and `lived` (with a small rounding error).

Exercise 3.61 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.[35]

Exercise 3.62 If 97.45% of the people who were inoculated lived, what proportion of inoculated people must have died?[36]

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:

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

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

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

Exercise 3.63 Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[37]

---

[35] The answer is 0.0382, which can be verified using Table 3.20.

[36] There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.45% = 2.55% of the people who were inoculated died.

[37] 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 Exercise (3.47), this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

3.2.6 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 3.22. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

Figure 3.22: A tree diagram of the `smallpox` data set.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 3.22. 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 3.22 is the probability that `result` = `lived` conditioned on the information that `inoculated` = `yes`. We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

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

Example 3.64 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 3.23. When constructing a tree diagram, variables provided with marginal probabilities are often used to create the tree's primary branches; in this case, the marginal probabilities are provided for midterm grades. The final grades, which correspond to the conditional probabilities provided, will be shown on the secondary branches.

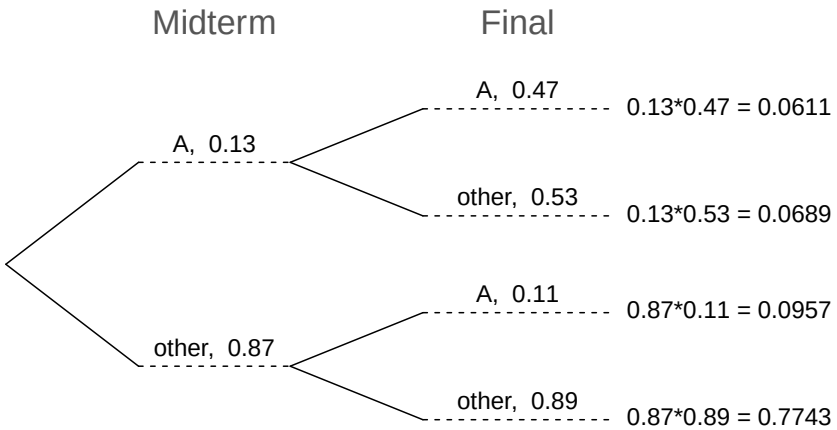


Figure 3.23: A tree diagram describing the `midterm` and `final` variables.

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

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

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

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

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

⊙ **Exercise 3.65** After an introductory statistics course, 78% of students can successfully construct tree diagrams. Of those who can construct tree diagrams, 97%

passed, while only 57% of those students who could not construct tree diagrams passed. (a) Organize this information into a tree diagram. (b) What is the probability that a randomly selected student passed? (c) Compute the probability a student is able to construct a tree diagram if it is known that she passed.[38]

3.2.7 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.

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:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)} \tag{3.66}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable. Bayes' Theorem is given more explicitly in the text box below.

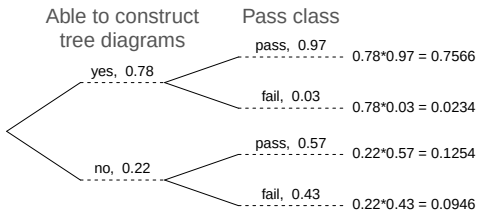
Bayes' Theorem: inverting probabilities

Let $A_1, A_2, \ldots, A_n$ be mutually exclusive events in sample space Ω such that $\sum_{i=1}^{n} P(A_i) = 1$. Let B be any event in Ω. Then:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{j=1}^{n} P(A_j) \cdot P(B|A_j)} \tag{3.67}$$

Bayes' Theorem is useful formula when manipulating conditional probabilities. Equation (3.66) means is that if we are provided information about $P(B|A_i)$, $P(B)$ and $P(A)$, we can recover $P(B_i|A)$.

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



Although (3.67) may look overwhelming it is quite intuitive once you become more familiar with its use. 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.

In many problems, it is usually sufficient to become familiar with the following forms of Baye's Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.68}$$

and

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)} \tag{3.69}$$

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

Example 3.70 In Canada, about 0.35% of women over 40 will be diagnosed with 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.[39] 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

[39] The probabilities reported here were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

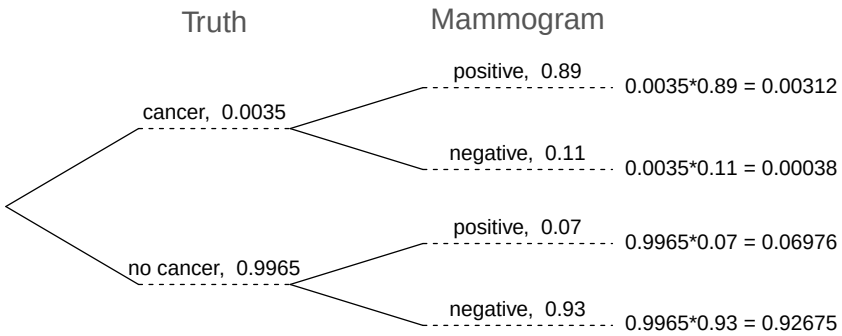


Figure 3.24: Tree diagram for Example (3.70), computing the probability a random patient who tests positive on a mammogram actually has breast cancer.

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 3.24. The probability the patient has breast cancer and the mammogram is positive is

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

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

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

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 (3.70) 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 (3.70). 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:

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

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.

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

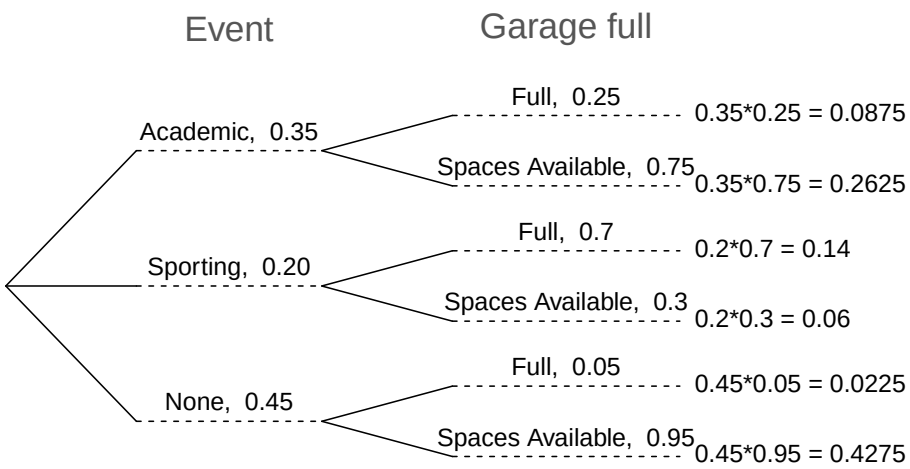
We can see an application of Bayes' Theorem by substituting the resulting probability expressions into the numerator and denominator of the original conditional probability.

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

Example 3.71

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

The tree diagram, with three primary branches, is shown to the right. 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.



Example 3.72 Here we solve the same problem presented in Example (3.71), 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 conditional 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.

Exercise 3.73 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.[40]

[40] Short answer:

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

⊙ **Exercise 3.74** In Example (3.71) and (3.73), you found that if the parking lot is full, the probability a sporting event is 0.56 and the probability there is an academic event is 0.35. Using this information, compute $P(\text{no event} \mid \text{the lot is full})$.[41]

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.

3.3 Sampling from a small population (special topic)

● **Example 3.75** Professors sometimes select a student at random to answer a question. If each student has an equal chance of being selected and there are 15 people in your class, what is the chance that she will pick you for the next question?

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

● **Example 3.76** 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

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

⊙ **Exercise 3.77** What rule permitted us to multiply the probabilities in Example (3.76)?[42]

● **Example 3.78** 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?

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

[42] 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.

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.

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

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.

⊙ **Exercise 3.79** Under the setup of Example (3.78), what is the probability of being picked to answer all three questions?[43]

If we sample from a small population **without replacement**, we no longer have independence between our observations. In Example (3.76), 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 (3.78), the professor sampled her students **with replacement**: she repeatedly sampled the entire class without regard to who she already picked.

⊙ **Exercise 3.80** 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?[44]

⊙ **Exercise 3.81** Compare your answers in Exercise (3.80). How much influence does the sampling method have on your chances of winning a prize?[45]

Had we repeated Exercise (3.80) 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.

3.4 Random variables

A **random variable** is the realization of what we previously called an experiment in which we obtain a numerical outcome. The term might be a little misleading since a random variable is actually a function.

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

[44] (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.

[45] 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.

Random variable
Let Ω be a sample space. A random variable X is a function that maps events in Ω to a real number.

$$X : \Omega \longrightarrow \mathbb{R} \tag{3.82}$$

The definition above is not a completely correct definition but it is sufficient for the purposes of an introductory course in statistics. We can consider a random variable to be a well defined map from our sample space to the real number line where each event in the sample space is mapped on to a some number. The proper definition involves a mathematical set of events known as a sigma algebra (or a sigma field). The definition involving sigma algebra is something that is more suited for a course in mathematical statistics or mathematical analysis (such as measure theory).[46]

We usually denote a random variable with a capital letter (such as "X") and the values it takes with a simple letter (such as x). The values that the random variable X can take is called the **support** of the random variable. Regarding notation, there are several well known probability distributions that a random variable can follow. Several of these distributions are discussed in Sections 4. If a random variable is follows a known (or well defined) distribution we usually denote this using the tilde ($\sim$) symbol.

Notation on distribution of a random variable
Let X be a random variable which follows a probability distribution $p(x)$. We say that X is distributed with distribution $p(x)$ and write this as

$$X \sim p(x) \tag{3.83}$$

There are 2 types of random variables which are discrete random variable and continuous random variable. Both of these are discussed in more detail in Sections 3.4.2 and 3.4.3.

3.4.1 Introduction of the mean and variance of random variables

At this point we would like to introduce the definition of mean and variance for random variables. These definitions will be repeated and used later but we introduce them in order to get familiar with the notation that we use as well to provide a gentle introduction to any subsequent courses in mathematical statistics. The box below is also a nice summary of the mean and variance of discrete and continuous random variables. The mean of a random variable is also called the **expected value**.

[46] For a thorough understanding, we should also be familiar with another mathematical object called an "algebra" because a sigma algebra is not an algebra.

Mean and variance of a random Variable

Let X be a random variable The mean of X is

$$E(X) = \mu = \begin{cases} \sum_{i=1}^{n} x_i \cdot P(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x \cdot f(x)\, dx & \text{if } X \text{ is continuous} \end{cases} \tag{3.84}$$

The variance of X, labeled σ^2, is

$$V(X) = \sigma^2 = E\big[\big(X - E(X)\big)^2\big] \tag{3.85}$$

$$= \begin{cases} \sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x)\, dx & \text{if } X \text{ is continuous} \end{cases} \tag{3.86}$$

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

$$SD(X) = \sigma = \sqrt{\sigma^2} \tag{3.87}$$

The mean μ of the random variable represents the average outcome. 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. In terms of a physical interpretation, the variance of a random variable can be considered as the moment of inertia around a vertical axis around the mean (which is the center of mass of the distribution).

Note that with some simple manipulation an easier way to compute the variance of a simple random variable is by using

$$V(X) = \sigma^2 = E(X)^2 - \mu^2 \tag{3.88}$$

It is possible to obtain (3.88) from (3.85) with some simple manipulation using properties of expectation.

TIP: Formulating the variance of a random variable

In many cases it is usually easier to calculate the variance of a random variable using (3.88) rather than (3.85). This is true for both discrete as we as continuous random variables.

⊙ **Exercise 3.89** Verify that (3.85) and (3.88) are equivalent forms of calculating the variance of a continuous random variable [Note: This may be a hard question for students in an introductory course].

3.4.2 Discrete Random variables

A **discrete random variable** is a random variable that can only take specific (i.e. discrete) values. Intuitively this means that we can list all of the possible values of the support of a discrete random variable. If the support of a discrete random variables is infinite, we can still list consecutive elements in an ordered list.

Example 3.90 Let X represent the value obtained by rolling a fair die. The possible values that X can take are $1, 2, 3, 4, 5$, or 6. X can not take anything between the values that we just mentioned. For instance, X can not be 2.5.

Example 3.91 Let X represent the number of customers that line up at a bank everyday. The possible values that X can take are $1, 2, 3, \ldots$. Theoretically these values can go on to infinity, however we can list consecutive elements in an ordered list. The bank can serve 99 customers or 100 customers but not something in between these two values.

3.4.2.1 Probability mass function

We can assign probabilities to the values that a random variable can take. We do this using a **probability distribution function**. For a discrete random variables the probability distribution function is called a **probability mass function (PMF)** or mass function for short. If X is discrete random variable which takes values $x = x_1, x_2, \ldots$ the probability of X taking on a specific value of x is written as $P(X = x_i)$ or $p(x_i)$ for short.

In order for a distribution function to be a probability mass function of a discrete random variable it must satisfy the following important properties:

Properties of a mass function

Let X be a discrete random variable which takes values $x = x_1, x_2, \ldots$ The probability mass function of X consists of individual probabilities that have the following properties:

1. $0 \leq p(x_i) \leq 1$
2. $\sum_{\forall i} p(x_i) = 1$

Property 1 is telling us that the probability for each possible outcome must be between 0 and 1. Property 2 is telling us that the sum of all of the probabilities should add up to 1. Note that we are being completely general in terms of the support of the random variable since its support can be finite or infinite.

3.4.2.2 Mean

Expected value of a discrete random variable

If X is a discrete random variable which takes the values of outcomes x_1, ..., x_n with probabilities $P(X = x_1)$, ..., $P(X = x_n)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$E(X) = \mu = \sum_{i=1}^{n} x_i \cdot P(X = x_i) \tag{3.92}$$

$$= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n) \tag{3.93}$$

It is important to make the distinction that $E(X)$ is not the same as the sample mean $\bar{x}$. We are calculating the actual mean of the process modelled by the discrete random variable.

Example 3.94 Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one term to another. If there are 100 students enrolled, how many books should the bookstore expect to sell to this class?

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

Exercise 3.95 Would you be surprised if the bookstore sold slightly more or less than 105 books?[47]

Example 3.96 The textbook costs \$137 and the study guide \$33. How much revenue should the bookstore expect from this class of 100 students?

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

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

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

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

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

[47] If they sell a little more or a little less, this should not be a surprise. Hopefully Chapter 2 helped make clear that there is natural variability in observed data. For example, if we would flip a coin 100 times, it will not usually come up heads exactly half the time, but it will probably be close.

Example 3.97 What is the average revenue per student for this course?

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

Example 3.98 Using the information in exercise (3.96), what is the probability mass function for the revenue per student for the course

The amount of money a single student will spend on her statistics books is a random variable, and we represent it by X. The possible outcomes of X are labeled with a corresponding lower case letter x and subscripts. We will write $x_1 = \$0$, $x_2 = \$137$, and $x_3 = \$170$, and these outcomes occur with probabilities 0.20, 0.55, and 0.25. The distribution of discrete random variable X (i.e. the probability mass function of X) is given in Table 3.25.

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

Table 3.25: The probability distribution for the random variable X, representing the bookstore's revenue from a single student.

Example 3.99 Using the solution to the problem in Example 3.4.2.2, plot the probability mass function of revenue per student for the course.

Using the mass function 3.25, we get point masses at each of the values in the support of X. The height that these point masses appear at will be their corresponds probabilities.

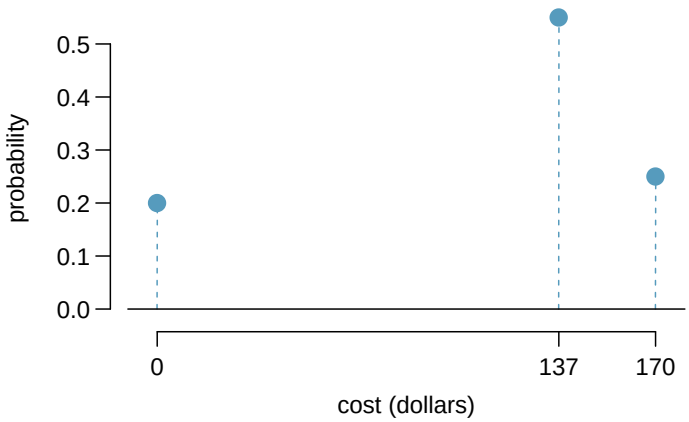


Figure 3.26: Plot of the probability mass function for the bookstore's revenue from a single student.

Example 3.100 Use the probability mass function from Example 3.4.2.2 and the definition of expected value for a discrete random variable given in equation (3.92) to calculate the mean revenue per student for this course

The expected value of a random variable is computed by adding each outcome weighted by its probability. Using (3.92) we get:

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

It is also possible to compute the expected value of a continuous random variable (see Section 3.4.2). However, it requires a little calculus and we save it for a later class.[48]

Recall the the explanation provided in Section 3.4.1 where we looked at the mean as the center of gravity. Figures 3.27 and 3.28 are representative of this explanation in terms of weighted amounts balanced on a pivot.

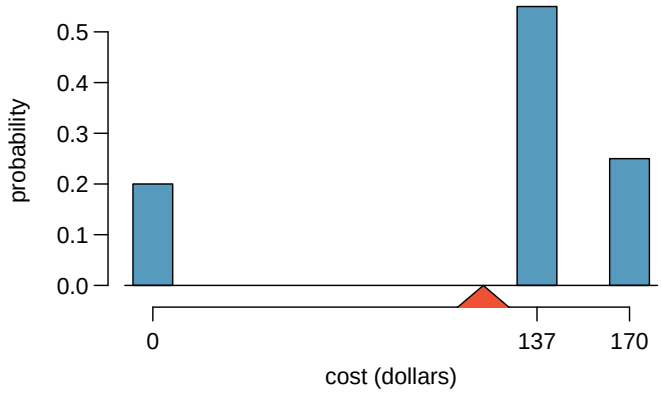


Figure 3.27: Probability distribution for the bookstore's revenue from a single student. The distribution balances on a triangle representing the average revenue per student.

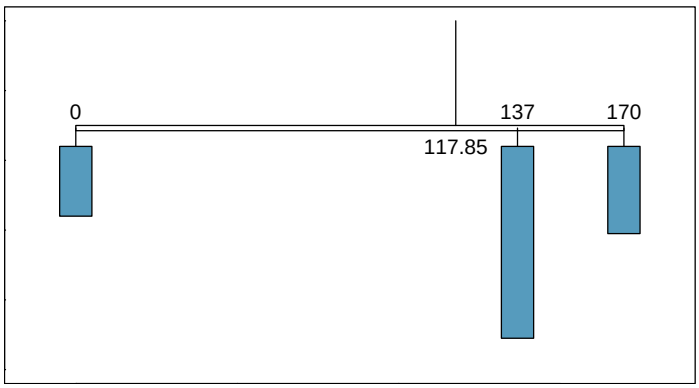


Figure 3.28: A weight system representing the probability distribution for X. The string holds the distribution at the mean to keep the system balanced.

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

3.4.2.3 Variance

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.

Variance of a discrete random variable

If X is a discrete random variable which takes the values of outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_n)$ and expected value $\mu = E(X)$, then the variance of X, denoted by $V(X)$ or the symbol σ^2, is

$V(X)$ Variance of X

$$V(X) = \sigma^2 = \sum_{i=1}^{k}(x_i - \mu)^2 \cdot P(X = x_i) \tag{3.101}$$

$$= (x_1 - \mu)^2 \cdot P(X = x_1) + \ldots + (x_n - \mu)^2 \cdot P(X = x_n) \tag{3.102}$$

The standard deviation of X is

$$SD(X) = \sigma = \sqrt{\sigma^2} \tag{3.103}$$

Example 3.104 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	
$x_i \cdot P(X = x_i)$	0	75.35	42.50	117.85

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

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \cdot P(X = x_i)$	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	
$(x_i - \mu)^2 \cdot P(X = x_i)$	2777.7	201.7	679.9	3659.3

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

⊙ **Exercise 3.105** 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.[49]

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

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

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

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

3.4.3 Continuous random variables

A **continuous random variable** is a random variable that can take any value within a specified range of values. The support of a continuous variable has infinite precision, meaning that we can make our measurement as fine as we would like. Recall that in Section 3.4.2 a discrete random variables was associated with a probability mass function.

We will now provide some motivation for the transition from a discrete distribution to a continuous distribution by introducing the concept of observing large amounts of data in smaller and smaller intervals.

● **Example 3.106** Figure 3.29 shows a few different hollow histograms of the variable `height` for 3 million US adults from the mid-90's.[50] How does changing the number of bins allow you to make different interpretations of the data?

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

[49] (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: $\sigma = \sqrt{V(Y)} = \$69.28$.

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

[50] This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

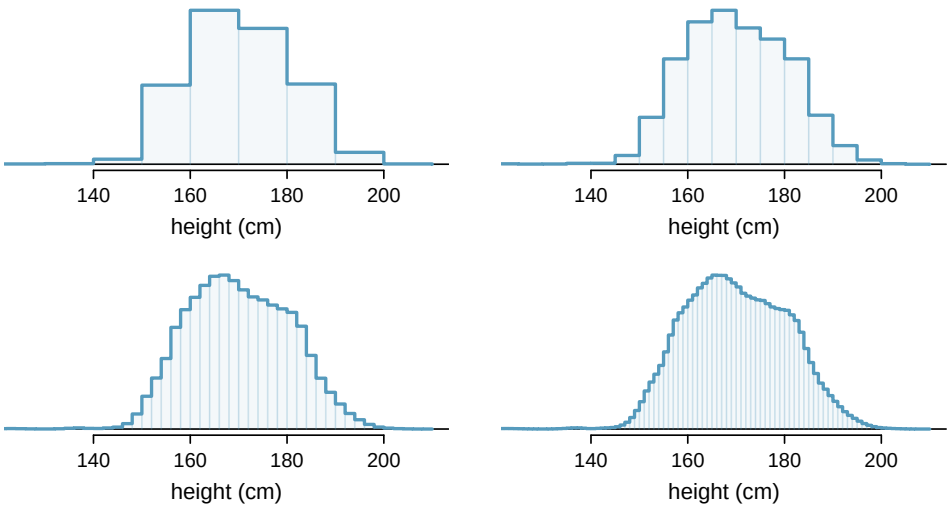


Figure 3.29: Four hollow histograms of US adults heights with varying bin widths.

Example 3.107 What proportion of the sample is between 180 cm and 185 cm tall (about 5'11" to 6'1")?

We can add up the heights of the bins in the range 180 cm and 185 and divide by the sample size. For instance, this can be done with the two shaded bins shown in Figure 3.30. 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.

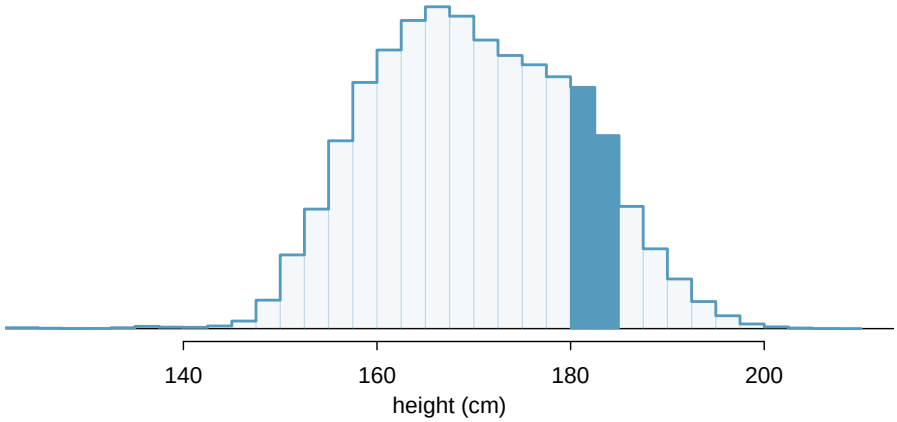


Figure 3.30: A histogram with bin sizes of 2.5 cm. The shaded region represents individuals with heights between 180 and 185 cm.

3.4.3.1 From histograms to continuous distributions

Examine the transition from a boxy hollow histogram in the top-left of Figure 3.29 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 3.31 overlaid on a histogram of the sample. A density has a special property: the total area under the density's curve is 1.

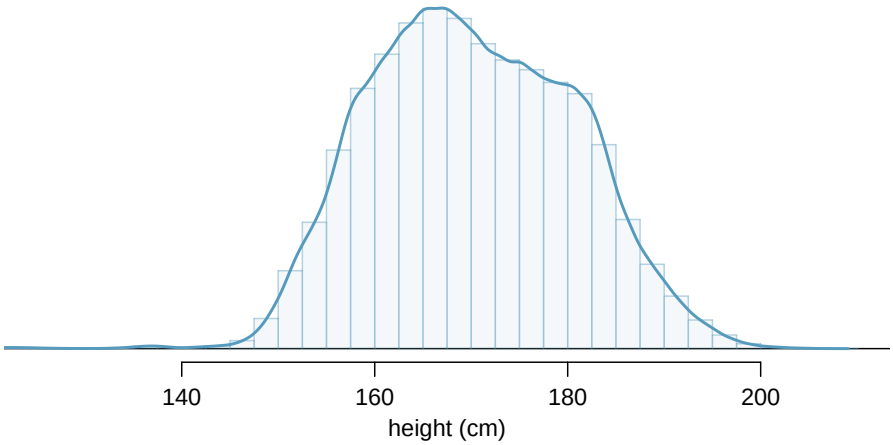


Figure 3.31: The continuous probability distribution of heights for US adults.

3.4.3.2 Probabilities from continuous distributions

We computed the proportion of individuals with heights `180` to `185` cm in Example (3.107) as a fraction:

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

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):

$$P(\texttt{height} \text{ between } \texttt{180} \text{ and } \texttt{185}) = \text{area between } \texttt{180} \text{ and } \texttt{185} = 0.1157$$

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 (3.107): 0.1172.

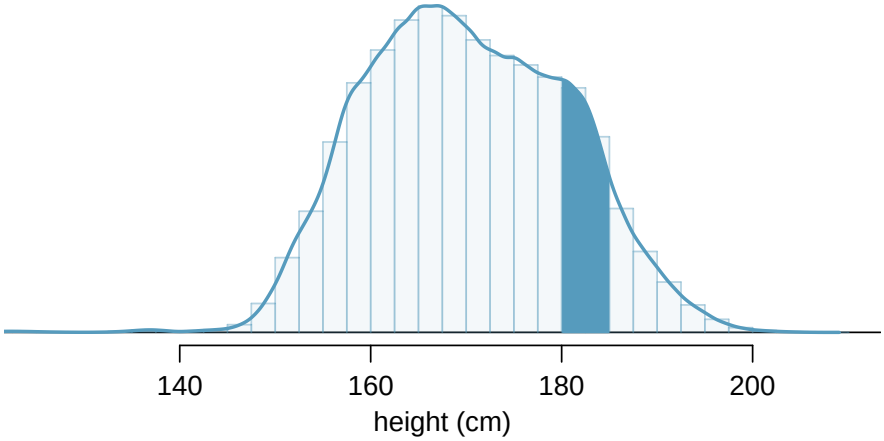


Figure 3.32: Density for heights in the US adult population with the area between 180 and 185 cm shaded. Compare this plot with Figure 3.30.

⊙ **Exercise 3.108** 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 3.109** 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.

⊙ **Exercise 3.110** 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]

3.4.3.3 Probability density function

We associate continuous random variables with a **probability density function (PDF)** or density function for short.

In order for a distribution function to be a density function of a continuous random variable it must satisfy the following important properties:

[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 (3.109).

Properties of a distribution function

Let X be a continuous random variable. We say that $f(x)$ is a probability mass function of X if it satisfies all of the following properties:

1. $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
2. $P(a \leq x \leq b) = \int_a^b f(x)\,dx$
3. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
4. $P(x = c) = 0, \quad \forall c \in \mathbb{R}$

A good knowledge of integral calculus is required in order to get a proper understanding of Properties 1, 2, 3 and 4. Integrals and derivatives are two of the fundamental concepts of calculus and the integral of a function gives a generalized notion of area (expressed by the fundamental theorem of calculus). We will explain 1 — 4 with minimal reference to calculus.

Recall in section 3.4.2 that when we plotted the mass function of a discrete random variable (such as in Example (3.99)) we got point masses at the values that the random variable takes. With a continuous random variables, the density function is a continuous curve where the domain is its support.

Property 1 implies that the curve of the density function must be above (or on) the horizontal axis. The density curve can never go below the horizontal axis. The definite integral of a function $f(x)$ from limits a to b gives the area between the curve and the x-axis (from calculus). Property 2 implies that for a continuous random variable, the probability of observing values between a and b is given by the area under the density curve between a and b. Since area under the curve gives probability, then the total area under the curve should give us the total probability (which is always 1). Property 3 implies that the total area under the curve should be equal to 1. In property 3 "$-\infty$" is notation for the lower bound of the support and "$+\infty$" is notation for the upper bound of the support. Finally property 3 implies that the probability of observing a single exact value c is 0. This is because the probability of observing a single exact value is given by the area under a line which is 0.

3.4.3.4 Mean

The calculation of the mean of a continuous random variable requires a knowledge of calculus.

Expected value of a continuous random variable

Let X be a continuous random variable with density function $f(x)$. Then

$$E(X) = \mu = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx \tag{3.111}$$

The calculation of the mean in (3.111) is the continuous analog to the calculation of the mean in (3.92) for the discrete case. Similar to property 3 in the list of properties of

a distribution function "$-\infty$" is notation for the lower bound of the support and "$+\infty$" is notation for the upper bound of the support.

Recall the concept of the weighted balancing point discussed in Section 3.4.1 We illustrated this concept in Figures 3.27 and 3.28 for the discrete case. The idea of a center of gravity also expands to continuous probability distributions. Figure 3.33 shows a continuous probability distribution balanced atop a wedge placed at the mean.

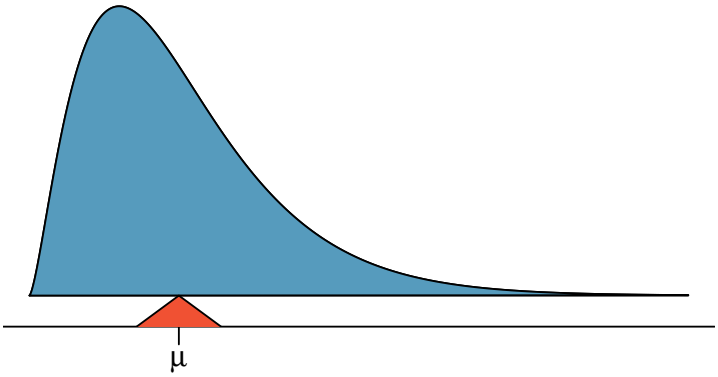


Figure 3.33: A continuous distribution can also be balanced at its mean.

3.4.3.5 Variance

The variance of continuous random variables is defined below.

Variance of a continuous random variable

Let X be a continuous random variable with density function $f(x)$. The variance of X is

$$V(X) = \sigma^2 = E[\big(X - E(X)\big)^2] \tag{3.112}$$

$$= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\, dx \tag{3.113}$$

The standard deviation of X is

$$SD(X) = \sigma = \sqrt{\sigma^2} \tag{3.114}$$

3.5 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.

A **linear combination** of two random variables is a fancy way os saying the following:

Linear combinations of random variables
If X and Y are random variables then a linear combination of X and Y is

$$aX + bY \tag{3.115}$$

where a and b are some fixed real numbers.

Example 3.116 John travels to work five days a week. We will use X_1 to represent his travel time on Monday, X_2 to represent his travel time on Tuesday, and so on. Write an equation using X_1, ..., X_5 that represents his travel time for the week, denoted by W.

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

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

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

Example 3.117 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.

She will make X dollars on the TV but spend Y dollars on the toaster oven, so the expression representing her net change in cash is

$$X - Y$$

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.

3.5.1 Expected value of linear combinations of random variables

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.

Expected Value of a linear combination of random variables
Let X and Y be random variables and let a and b be any two real numbers. To compute the average value of a linear combination of X and Y, plug in the average of each individual random variable and compute the result:

$$E(aX + bY) = a \cdot E(X) + b \cdot E(Y)$$

Example 3.118 Leonard has invested \$6000 in Google Inc. (stock ticker: GOOG) and \$2000 in Exxon Mobil Corp. (XOM). If X represents the change in Google's stock next month and Y represents the change in Exxon Mobil stock next month, write an equation that describes how much money will be made or lost in Leonard's stocks for the month.

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

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

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

Example 3.119 It takes John an average of 18 minutes each day to commute to work. What would you expect his average commute time to be for the week?

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

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

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

Exercise 3.120 Suppose Google and Exxon Mobil stocks have recently been rising 2.1% and 0.4% per month, respectively. Compute the expected change in Leonard's stock portfolio for next month.[53]

Exercise 3.121 You should have found that Leonard expects a positive gain in Exercise (3.120). However, would you be surprised if he actually had a loss this month?[54]

Exercise 3.122 Recall Example (3.117). 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?[55]

Exercise 3.123 With reference Example (3.119) and the solution to Exercise (3.122), would you be surprised if John's weekly commute wasn't exactly 90 minutes or if Elena didn't make exactly \$152? Explain.[56]

---

[53] $E(\$6000 \times X + \$2000 \times Y) = \$6000 \times 0.021 + \$2000 \times 0.004 = \$134$.

[54] No. While stocks tend to rise over time, they are often volatile in the short term.

[55] $E(X - Y) = E(X) - E(Y) = 175 - 23 = \$152$. She should expect to make about \$152.

[56] 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.

For a few examples of nonlinear combinations of random variables – cases where we cannot simply plug in the means – see the footnote.[57]

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*.

3.5.2 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 Exercise (3.120). However, there was no quantitative discussion of the volatility of this portfolio. For instance, while the average monthly gain might be about $134 according to the data, that gain is not guaranteed. Figure 3.34 shows the monthly changes in a portfolio like Leonard's during the 36 months from 2009 to 2011. The gains and losses vary widely, and quantifying these fluctuations is important when investing in stocks.

Figure 3.34: The change in a portfolio like Leonard's for the 36 months from 2009 to 2011, where $6000 is in Google's stock and $2000 is in Exxon Mobil's.

Just as we have done in many previous cases, we use the variance and standard deviation to describe the uncertainty associated with Leonard's monthly returns. To do so, the variances of each stock's monthly return will be useful, and these are shown in Table 3.35. The stocks' returns are nearly independent.

	Mean ($\bar{x}$)	Standard deviation (s)	Variance (s^2)
GOOG	0.0210	0.0846	0.0072
XOM	0.0038	0.0519	0.0027

Table 3.35: The mean, standard deviation, and variance of the GOOG and XOM stocks. These statistics were estimated from historical stock data, so notation used for sample statistics has been used.

[57] 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.

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 (i.e. The variance of a linear combination of random variables can be computed by squaring the constants, substituting in the variances for the random variables, and calculating the result).

Variability of linear combinations of random variables

Let X and Y be random variables. The variance of a linear combination of X and Y is given by

$$V(aX + bY) = a^2V(X) + b^2V(Y) \tag{3.124}$$

where a and b are fixed real numbers.

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.

We reiterate the importance of assuming that the random variables are independent in (3.124). If independence doesn't hold, then more advanced methods are necessary. Equation (3.124) can be used to compute the variance of Leonard's monthly return:

$$\begin{aligned} V(6000X + 2000Y) &= 6000^2 \times V(X) + 2000^2 \times V(Y) \\ &= 36,000,000 \times 0.0072 + 4,000,000 \times 0.0027 \\ &= 270,000 \end{aligned}$$

The standard deviation is computed as the square root of the variance. In the case of Leonard's monthly return

$$\begin{aligned} SD(6000X + 2000Y) &= \sqrt{V(6000X + 2000Y)} \\ &= \sqrt{270,000} = \$520 \end{aligned}$$

While an average monthly return of \$134 on an \$8000 investment is nothing to scoff at, the monthly returns are so volatile that Leonard should not expect this income to be very stable.

Example 3.125 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.

⊙ **Exercise 3.126** The computation in Example (3.125) 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.[58]

⊙ **Exercise 3.127** Consider Elena's two auctions from Exercise (3.117) on page 96. 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.[59]

Consider again Exercise (3.127). 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.

[58] 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.

[59] 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:

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

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

Chapter 4

Distributions of random variables

4.1 Distributions of discrete random variables

4.1.1 Bernoulli distribution

Bernoulli random variable is a discrete random variable that has exactly two possible outcomes which are either a `success` or a `failure`. An experiment in which there are exactly 2 outcomes (which are success or failure) is called a **Bernoulli trial**. The mass function of a Bernoulli random variable is given in the box below.

Bernoulli distribution
Let $X \sim \text{Bernoulli}(p)$. This means that X is a random variable that takes value 1 with probability of success p and 0 with probability $1-p$. The mass function of X is

$$P(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1 \tag{4.1}$$

When $x = 1$ we have a success and when $x = 0$ we have a failure. The term success and failure are relative to the problem being studied.

TIP: "success" need not be something positive
We chose to label a person who refuses to administer the worst shock 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.

Example 4.2 A jar of marbles has 10 red marbles and 6 blue marbles. We consider drawing a red marble as a success. What is the probability mass function?

There are a total of 16 marbles in the jar. Therefore the probability of drawing a red

marble is $p = \frac{10}{16} = 0.625$. Let $X \sim \text{Bernoulli}(0.625)$. The mass function of X is

$$P(X = x) = 0.625^x(0.385)^{1-x}, \quad x = 0, 1 \tag{4.3}$$

To illustrate consider the famous psychology experiment conducted by Milgram. Stanley Milgram began a series of experiments in 1963 to estimate what proportion of people would willingly obey an authority and give severe shocks to a stranger. Milgram found that about 65% of people would obey the authority and give such shocks. Over the years, additional research suggested this number is approximately consistent across communities and time.[1]

Each person in Milgram's experiment can be thought of as a Bernoulli trial. We label a person a success if they refuse to administer the worst shock. A person is labeled a failure if they administer the worst shock. Since only 35% of individuals refused to administer the most severe shock, we denote the probability of a success with $p = 0.35$. The probability in this experiment is $1 - p = 0.65$. So success or failure was recorded for each person in Milgram's study. Administering the test on each person is a example of a Bernoulli trial since each individual trial has only two possible outcomes (success of failure) and is therefore a Bernoulli random variable

Since 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:

```
0 1 1 1 1 0 1 1 0 0
```

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\#\text{ of successes}}{\#\text{ of trials}} = \frac{0+1+1+1+1+0+1+1+0+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. 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.

Since `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable. It turns out that a Bernoulli random variable has a nice characterization. It turns out that random variables that follow many other other statistical distributions also have nice closed forms.

Mean and variance of a Bernoulli random variable

Let $X \sim \text{Bernoulli}(p)$. The mean of X is

$$E(X) = \mu = p \tag{4.4}$$

and the variance of X is

$$V(X) = \sigma^2 = p(1-p) \tag{4.5}$$

We can obtain (4.4) and (4.5) by using (3.92) and (3.101). Although the derivations of the mean and standard deviations of random variables are more suited for a course in mathematical statistics we will show how we obtained (3.92) and (3.101).

[1] Find further information on Milgram's experiment at www.cnr.berkeley.edu/ucce50/ag-labor/7article/article35.htm.

Let X be a Bernoulli random variable with the probability of a success as p. Then

$$E[X] = \mu = \sum_{i=1}^{n} x_i \cdot P(X = x_i) \qquad \text{(recall (3.92))} \tag{4.6}$$
$$= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \tag{4.7}$$
$$= 0 \cdot (1 - p) + 1 \cdot p \tag{4.8}$$
$$= p \tag{4.9}$$

Similarly, the variance of X can be computed:

$$V(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i) \qquad \text{(recall (3.101))} \tag{4.10}$$
$$= (0 - p)^2 \cdot P(X = 0) + (1 - p)^2 \cdot P(X = 1) \tag{4.11}$$
$$= p^2(1 - p) + (1 - p)^2 p \tag{4.12}$$
$$= p(1 - p) \tag{4.13}$$

The standard deviation is

$$\sigma = \sqrt{\sigma^2} \tag{4.14}$$
$$= \sqrt{p(1 - p)} \tag{4.15}$$

4.1.2 Binomial distribution

In section 4.1.1 we learnt about Bernoulli random variables in which we were interested in the outcome of just a single trial. A **binomial random variable** is a generalization of several independent Bernoulli trials. Instead of performing just a single Bernoulli trial and observing whether we have a success or not, we are now performing several Bernoulli trials and observing whether we have a certain number of successes and failures. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p.

Binomial distribution

Let $X \sim \text{Bin}(n, p)$. This means that the probability of a single trial being a success is p and we conduct n such independent trials. The probability of observing exactly k successes in these n independent trials is given by

$$P(X = k) \quad = \quad \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n \tag{4.16}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n - k)!} \tag{4.17}$$

We will explain (4.16) in a bit more detail so that it makes more sense. We will illustrate with Example (4.18)

Example 4.18 Suppose we randomly selected four individuals to participate in the "shock" study. What is the chance exactly one of them will be a success? Let's call the four people Allen (A), Brittany (B), Caroline (C), and Damian (D) for convenience. Also, suppose 35% of people are successes as in the previous version of this example.

Let's consider a scenario where one person refuses:

$$\begin{aligned} &P(A = \texttt{refuse},\ B = \texttt{shock},\ C = \texttt{shock},\ D = \texttt{shock}) \\ &\quad = P(A = \texttt{refuse})\ P(B = \texttt{shock})\ P(C = \texttt{shock})\ P(D = \texttt{shock}) \\ &\quad = (0.35)(0.65)(0.65)(0.65) = (0.35)^1(0.65)^3 = 0.096 \end{aligned}$$

But there are three other scenarios: Brittany, Caroline, or Damian could have been the one to refuse. In each of these cases, the probability is again $(0.35)^1(0.65)^3$. These four scenarios exhaust all the possible ways that exactly one of these four people could refuse to administer the most severe shock, so the total probability is $4 \times (0.35)^1(0.65)^3 = 0.38$.

Exercise 4.19 Verify that the scenario where Brittany is the only one to refuse to give the most severe shock has probability $(0.35)^1(0.65)^3$.[2]

The scenario outlined in Example (4.18) is a special case of the binomial distribution (in Example (4.18), $n = 4$, $k = 1$, $p = 0.35$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want to identify n, k, and p to obtain the desired probability. To do this, we reexamine each part of the example.

There were four individuals who could have been the one to refuse, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\text{final probability}] \quad = \quad [\#\text{ of scenarios}] \times P(\text{single scenario}) \tag{4.20}$$

The first component of (4.16) is the number of ways to arrange the $k = 1$ 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:

$$p^k(1-p)^{n-k} \tag{4.21}$$

This is our general formula for $P(\text{single scenario})$.

Secondly, we explain a general formula for the number of ways to choose k successes in n trials, i.e. arrange k successes and $n - k$ failures:

$$\binom{n}{k} \quad = \quad \frac{n!}{k!(n-k)!} \tag{4.22}$$

The quantity $\binom{n}{k}$ is read **n choose k**.[3] The exclamation point notation (e.g. $k!$) denotes

[2] $P(A = \texttt{shock},\ B = \texttt{refuse},\ C = \texttt{shock},\ D = \texttt{shock}) = (0.65)(0.35)(0.65)(0.65) = (0.35)^1(0.65)^3$.

[3] Other notation for n choose k includes ${}_nC_k$, C_n^k, and $C(n, k)$.

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 $k = 1$ successes in $n = 4$ trials:

$$\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4!}{1!3!} = \frac{4 \times 3 \times 2 \times 1}{(1)(3 \times 2 \times 1)} = 4$$

This result is exactly what we found by carefully thinking of each possible scenario in Example (4.18).

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability in Equation ((4.20)). Therefore our final answer using (4.16) is

$$\begin{aligned}
P(X = 1) &= \binom{4}{1} 0.35^1 (1 - 0.35)^{4-1} \\
&= 4 \times 0.35 \times (0.65)^3 \\
&= 0.384475
\end{aligned}$$

We should note that in the binomial distribution the probability of a success on any trial stays fixed at p. By this we mean that the probability of a success on the first trial is p and the probability of a success on the second trial is still p and so on. This is true because we assume that the trials are independent from one another. Since p represents the probability of a success on any trial this means that $1 - p$ is the probability of a failure on any trial.

In (4.16) n must be a positive integer. If $n = 1$ then we go back to the case of a Bernoulli trial. The value of k should be a non-negative integer. Since k is the number of successes, the smallest value that k can be is 0 (i.e. no successes in all n trials) and the largest number that k can be is n (i.e. all trials are successes). If we have k successes in n trials the remaining $n - k$ trials must be failures.

TIP: Is it binomial? Four conditions to check.

1. The trials are independent.
2. The number of trials, n, is fixed.
3. Each trial outcome can be classified as a *success* or *failure*.
4. The probability of a success, p, is the same for each trial.

Example 4.23 What is the probability that 3 of 8 randomly selected students will refuse to administer the worst shock (i.e. 5 of 8 will administer the worst shock)?

We would like to apply the binomial model, so we check our 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 $k = 3$ successes in $n = 8$ trials, and the probability of a success is $p = 0.35$. So the probability that 3 of 8 will refuse is given by

$$\begin{aligned}\binom{8}{3}(0.35)^3(1-0.35)^{8-3} &= \frac{8!}{3!(8-3)!}(0.35)^3(1-0.35)^{8-3} \\ &= \frac{8!}{3!5!}(0.35)^3(0.65)^5\end{aligned}$$

Dealing with the factorial part:

$$\frac{8!}{3!5!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(3\times2\times1)(5\times4\times3\times2\times1)} = \frac{8\times7\times6}{3\times2\times1} = 56$$

Using $(0.35)^3(0.65)^5 \approx 0.005$, the final probability is about $56 \times 0.005 = 0.28$.

TIP: 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. The final step is to apply the formulas and interpret the results.

TIP: computing n choose k
In general, it is useful to do some cancelation in the factorials immediately. Alternatively, many computer programs and calculators have built in functions to compute n choose k, factorials, and even entire binomial probabilities.

⊙ **Exercise 4.24** 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?[4]

⊙ **Exercise 4.25** 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 (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?[5]

[4] 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.

[5] 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) from the binomial formula in Equation ((4.16)): $P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1\times1\times0.7^4 = 0.2401$, $P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116$. Note: $0! = 1$, as shown on page 105. 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.

⊙ **Exercise 4.26** What is the probability that at least 2 of your 4 smoking friends will develop a severe lung condition in their lifetimes?[6]

⊙ **Exercise 4.27** Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers. (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.[7]

The mean and variance of a binomial random variable are given below

Mean and variance of a binomial random variable
Let $X \sim Bin(n, p)$. The mean of X is

$$E(X) = \mu = np \tag{4.28}$$

and the variance of X is

$$V(X) = \sigma^2 = np(1-p) \tag{4.29}$$

● **Example 4.30** If you ran a study and randomly sampled 40 students, how many would you expect to refuse to administer the worst shock? What is the standard deviation of the number of people who would refuse? Equation ((4.28)) and ((4.29)) may be useful.

We are asked to determine the expected number (the mean) and the standard deviation, both of which can be computed from the formulas in Equation ((4.28)) and ((4.29)):

$$\begin{aligned} \mu &= np \\ &= 40 \times 0.35 \\ &= 14 \end{aligned}$$

and

$$\begin{aligned} \sigma &= \sqrt{np(1-p)} \\ &= \sqrt{40 \times 0.35 \times 0.65} \\ &= 3.02 \end{aligned}$$

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 8 but less than 20 individuals in our sample who would refuse to administer the shock.

Below we consider the first term in the binomial probability, n choose k under some special scenarios.

[6] The complement (no more than one will develop a severe lung condition) as computed in Exercise (4.25) as 0.6517, so we compute one minus this value: 0.3483.

[7] (a) $\mu = 0.3 \times 7 = 2.1$. (b) $P(0, 1, \text{or } 2 \text{ develop severe lung condition}) = P(k = 0) + P(k = 1) + P(k = 2) = 0.6471$.

Example 4.31 Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any non-negative integer n?

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.)

Exercise 4.32 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?[8]

4.1.3 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.

Example 4.33 Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict the worst shock.[9] If the probability a person will *not* give the most severe shock is still 0.35 and the subjects are independent, what are the chances that she will stop the study after the first person? The second person? The third? What about if it takes her $n-1$ individuals who will administer the worst shock before finding her first success, i.e. the first success is on the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

The probability of stopping after the first person is just the chance the first person will not administer the worst shock: $1 - 0.65 = 0.35$. The probability it will be the second person is

$$\begin{aligned} &P(\text{second person is the first to not administer the worst shock}) \\ &\quad = P(\text{the first will, the second won't}) = (0.65)(0.35) = 0.228 \end{aligned}$$

Likewise, the probability it will be the third person is $(0.65)(0.65)(0.35) = 0.148$.

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 $(0.65)^{n-1}(0.35)$. This is the same as $(1-0.35)^{n-1}(0.35)$.

Example (4.33) illustrates what is called 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

[8] 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:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

[9] This is hypothetical since, in reality, this sort of study probably would not be permitted any longer under current ethical standards.

in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example (4.33) is shown in Figure 4.1. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

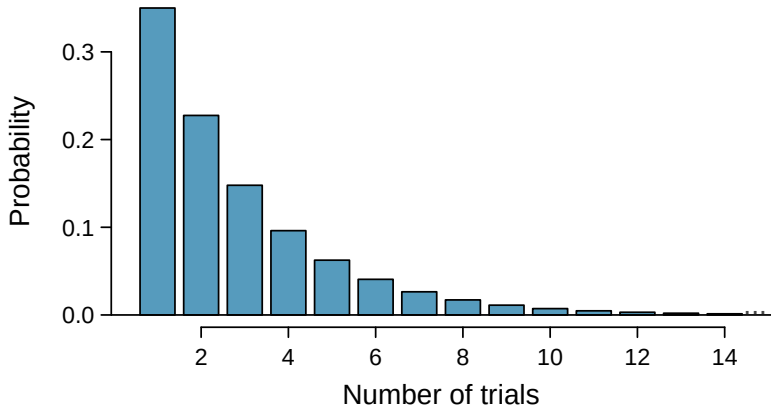


Figure 4.1: The geometric distribution when the probability of success is $p = 0.35$.

Geometric Distribution

Let $X \sim \text{Geom}(p)$. This means that the probability of a success in one trial is p and the probability of a failure is $1-p$ so the probability of finding the first success in the n^{th} trial (when trials are independent) is given by

$$(1-p)^{n-1}p, \quad n = 1, 2, 3, \ldots \tag{4.34}$$

A more in intuitive way to look at (4.34) is the following:

$$\begin{aligned} P(\text{success on n}^{th}\text{ trial}) &= \underbrace{P(\text{failure}) \cdot P(\text{failure}) \cdot \ldots \cdot P(\text{failure})}_{n-1 \text{ times}} \cdot P(\text{success}) \\ &= \underbrace{(1-p)(1-p)\ldots(1-p)}_{n-1 \text{ times}} p \\ &= (1-p)^{n-1}p \end{aligned}$$

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. The derivation of the mean and the variance of a geometric random variable requires a knowlede on mathematical series. In particular we require knowledge of the geometric series (from which this distribution gets its name) where consecutive terms differ by a common ratio with modulus less than 1; as well as a knowledge of the sum to infinity of a geometric series series.

Mean and variance of a geometric random variable
Let $X \sim Geom(p)$. The mean of X is

$$E(X) = \mu = \frac{1}{p} \tag{4.35}$$

and the variance of X is

$$V(X) = \sigma^2 = \frac{1-p}{p^2} \tag{4.36}$$

Equation ((4.35)) says that, on average, it takes $1/p$ trials to get a success. 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.

⊙ **Exercise 4.37** The probability that an individual would refuse to administer the worst shock is said to be about 0.35. If we were to examine individuals until we found one that did not administer the shock, how many people should we expect to check? Equation ((4.35)) may be useful.[10]

● **Example 4.38** What is the chance that Dr. Smith will find the first success within the first 4 people?

This is the chance it is the first ($n = 1$), second ($n = 2$), third ($n = 3$), or fourth ($n = 4$) person as the first success, which are four 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:

$$\begin{aligned} &P(n = 1, 2, 3, \text{ or } 4) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) + P(n = 4) \\ &\quad = (0.65)^{1-1}(0.35) + (0.65)^{2-1}(0.35) + (0.65)^{3-1}(0.35) + (0.65)^{4-1}(0.35) \\ &\quad = 0.82 \end{aligned}$$

There is an 82% chance that she will end the study within 4 people.

⊙ **Exercise 4.39** Determine a more clever way to solve Example (4.38). Show that you get the same result.[11]

● **Example 4.40** Suppose in one region it was found that the proportion of people who would administer the worst shock was "only" 55%. If people were randomly selected from this region, what is the expected number of people who must be checked before one was found that would be deemed a success? What is the standard deviation of this waiting time?

A success is when someone will **not** inflict the worst shock, which has probability $p = 1 - 0.55 = 0.45$ for this region. The expected number of people to be checked is $1/p = 1/0.45 = 2.22$ and the standard deviation is $\sqrt{(1-p)/p^2} = 1.65$.

[10] We would expect to see about $1/0.35 = 2.86$ individuals to find the first success.

[11] First find the probability of the complement: $P(\text{no success in first 4 trials}) = 0.65^4 = 0.18$. Next, compute one minus this probability: $1 - P(\text{no success in 4 trials}) = 1 - 0.18 = 0.82$.

⊙ **Exercise 4.41** Using the results from Example (4.40), $\mu = 2.22$ and $\sigma = 1.65$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[12]

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.

4.1.4 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?

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.[13]

TIP: 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.

⊙ **Exercise 4.43** 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.[14]

● **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.

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

[12] No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.

[13] See a similar guide for the binomial distribution on page 105.

[14] 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.

sequences of these first five kicks, which are shown in Table 4.2. 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.

	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	$\overset{4}{S}$

Table 4.2: The ten possible sequences when the fourth successful kick is on the sixth attempt.

⊙ **Exercise 4.45** Each sequence in Table 4.2 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.[15]

If the probability Brian kicks a 35 yard field goal is $p = 0.8$, what is the probability it takes 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 Table 4.2. 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 Exercise (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 4.1.2. The probability is broken into two pieces:

$$\begin{aligned} &P(\text{it takes Brian six tries to make four field goals}) \\ &= [\text{Number of possible sequences}] \times P(\text{Single sequence}) \end{aligned}$$

[15] The first sequence: $0.2 \times 0.2 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.0164$.

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 105.

Negative binomial distribution

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1}p^k(1-p)^{n-k} \qquad (4.46)$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

Example 4.47 Show using Equation ((4.46)) that the probability Brian kicks his fourth successful field goal on the sixth attempt is 0.164.

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 $n = 6$.

$$\binom{n-1}{k-1}p^k(1-p)^{n-k} = \frac{5!}{3!2!}(0.8)^4(0.2)^2 = 10 \times 0.0164 = 0.164$$

Exercise 4.48 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?[16]

Exercise 4.49 Assume Brian's kick attempts are independent. What is the probability that Brian will kick his fourth field goal within 5 attempts?[17]

[16] 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.

[17] If his fourth field goal ($k = 4$) is within five attempts, it either took him four or five tries ($n = 4$ or

TIP: 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.

⊙ **Exercise 4.50** 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.[18]

(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?

4.1.5 Poisson distribution

● **Example 4.51** 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?

A histogram of the number of occurrences of AMI on 365 days[19] for NYC is shown in Figure 4.3. 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.

The **Poisson distribution** is often useful for estimating the number of rare events in a large population over a unit of time. For instance, consider each of the following events, which are rare for any given individual:

- having a heart attack,
- getting married, and

$n = 5$). We have $p = 0.8$ from earlier. Use Equation ((4.46)) to compute the probability of $n = 4$ tries and $n = 5$ tries, then add those probabilities together:

$$P(n = 4 \text{ OR } n = 5) = P(n = 4) + P(n = 5)$$
$$= \binom{4-1}{4-1} 0.8^4 + \binom{5-1}{4-1} (0.8)^4 (1 - 0.8) = 1 \times 0.41 + 4 \times 0.082 = 0.41 + 0.33 = 0.74$$

[18] 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.

[19] These data are simulated. In practice, we should check for an association between successive days.

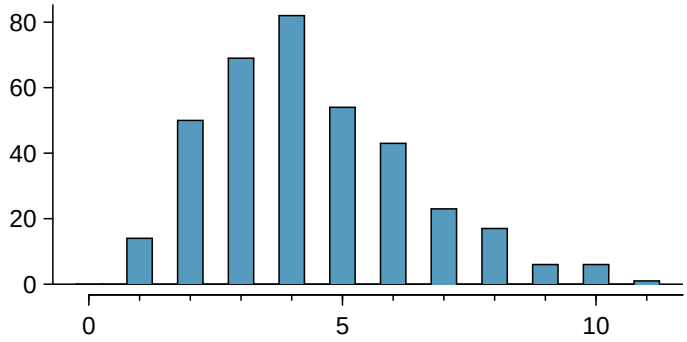


Figure 4.3: A histogram of the number of occurrences of AMI on 365 separate days in NYC.

- getting struck by lightning.

The Poisson distribution helps us describe the number of such events that will occur in a short unit of time for a fixed population if the individuals within the population are independent.

The histogram in Figure 4.3 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 (4.51), 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 rare 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 rare events in a single unit of time.

λ
Rate for the Poisson dist.

Poisson distribution
Suppose we are watching for rare events and the number of observed events follows a Poisson distribution with rate λ. Then

$$P(\text{observe } k \text{ rare events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page 105. 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.

TIP: Is it Poisson?
A random variable may follow a Poisson distribution if the event being considered is rare, the population is large, and the events occur independently of each other.

Even when rare events are not really independent – for instance, Saturdays and Sundays are especially popular for weddings – a Poisson model may sometimes still be reason-

able 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 **??** and **??**, we will discuss a foundation of linear models.

4.2 Distributions of continuous random variables

4.2.1 Continuous uniform distribution

A continuous random variable which has an equally likely chance of taking any value within an defined interval $[a, b]$ follows a **uniform probability distribution**. For example there may be an equally likely chance that a health and safety inspector may appear during regular business hours at a particular industrial plant. The continuous uniform distribution serves as a basic but important foundation for later courses.

The density function of a continuous random variable is very simple.

Continuous uniform distribution
Let $X \sim \mathrm{U}(a, b)$. This means that there is an equally likely chance that an outcome will occur over the interval $[a, b]$. The density function of X is

$$f(x) = \begin{cases} \dfrac{1}{(b-a)} & \text{for } \ a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The continuous uniform distribution takes has a constant height of $\frac{1}{b-a}$ over horizontal axis on the interval $[a, b]$. Recall from Property 3 of the properties of a distribution function that we calculate the probability of observing an outcome between two values by calculating the integral of the density function between these two values. We could use integral calculus to calculate probabilities but in the case of the continuous uniform distribution we can simply find the area of the rectangle between our two values of interest. This is very since the area of a rectangle is (base)×(height).

Example 4.52 Preparing a cup of cappuccino involves several steps such as grinding beans, making the espresso and steaming milk. There is an equally likely chance that a particular barista takes between 30 seconds and 50 seconds to make a cup of cappuccino (based on his experience and the equipment used). Suppose that a customer goes to order a cappuccino from this particular barista. What is the probability distribution function describing the amount of time taken by the barista to make a cup of cappuccino?

Let X be a random variable that represent the amount of time taken by this barista to make a cup of cappuccino. Since there is an equally likely chance of taking between 30 seconds and 50 seconds, this means that the time taken to make a cup of cappuccino is uniformly distributed between 30 seconds and 50 seconds. i.e. $X \sim U(30, 50)$. The height of the density curve is the constant value of $\frac{1}{30-50} = \frac{1}{20} = 0.05$ The distribution

function of X is

$$f(x) = \begin{cases} 0.05 & \text{for } \ 30 \leq x \leq 50 \\ \\ 0 & \text{otherwise} \end{cases}$$

Example 4.53 Plot the density function in Example (4.52)

Using the density function in Example (4.52) we obtain the following plot

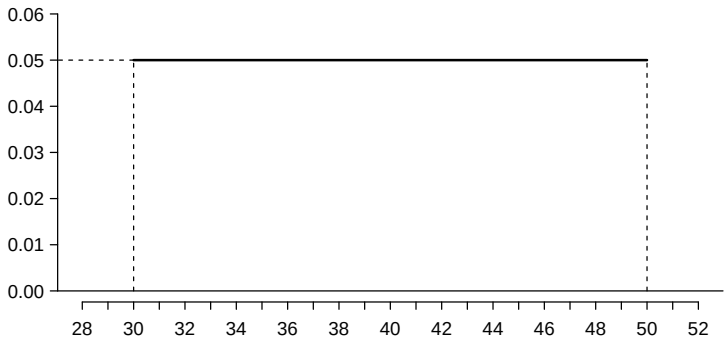


Figure 4.4: Plot of the continuous uniform distribution for Example (4.52)

Example 4.54 What is the probability that the barista will take less than 45 seconds to finish making a cup of cappuccino for this customer?

Recall that the area under the density curve gives us the probability of interest. Let's shade in the probability of the barista finishing under 45 seconds.

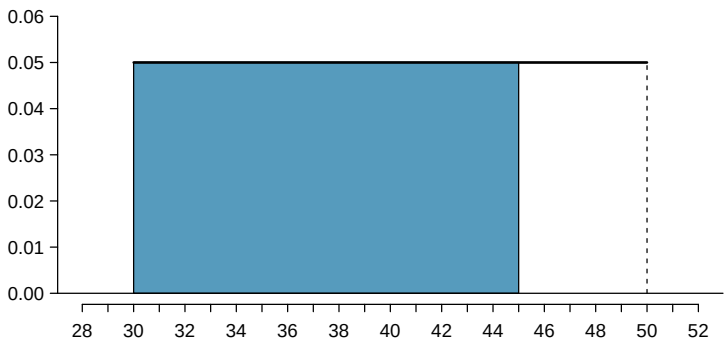


Figure 4.5: The shaded area gives $P(x \leq 45)$

We can calculate this area using integral calculus.

$$\begin{aligned} P(x \leq 45) &= \int_{30}^{45} 0.05\, dx \\ &= \big[0.05x\big]_{30}^{45} \\ &= 0.05(45) - 0.05(30) \\ &= 0.75 \end{aligned}$$

However since the shaded area in Figure 4.5 is a rectangle we can calculate $P(x \leq 45)$ using the formula for the area of a rectangle.

$$\begin{aligned} P(x \leq 45) &= (\text{base})(\text{height}) \\ &= (45 - 30)(0.05) \\ &= 0.75 \end{aligned}$$

Example 4.55 What is the probability that the barista will take between 35 seconds and 40 seconds to finish making a cup of cappuccino for a customer?

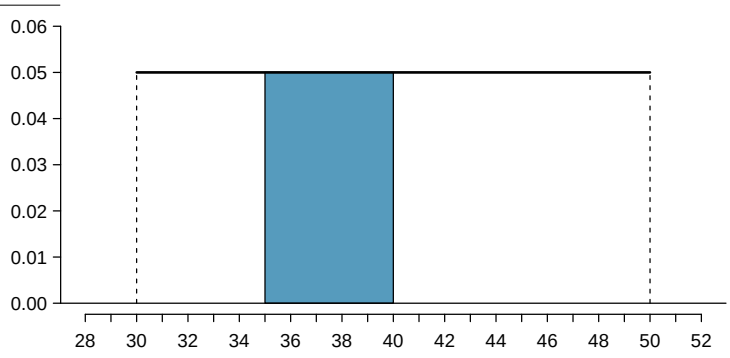


Figure 4.6: The shaded area gives $P(x \leq 45)$

By using integral calculus we get

$$\begin{aligned} P(35 \leq x \leq 45) &= \int_{30}^{45} 0.05\, dx \\ &= \big[0.05x\big]_{35}^{45} \\ &= 0.05(45) - 0.05(35) \\ &= 0.5 \end{aligned}$$

Using the formula for the area of a rectangle

$$\begin{aligned} P(x \leq 45) &= (\text{base})(\text{height}) \\ &= (45 - 35)(0.05) \\ &= 0.50 \end{aligned}$$

⊙ **Exercise 4.56** Given that the barista takes more than 40 seconds, what is the probability that he takes less than 45 seconds? [20]

● **Example 4.57** Let t be the time such that 80% of all cups of cappuccino made by this barista take less that t seconds (i.e. t is the 80^{th} percentile in terms of time taken to make a cup of cappuccino). Calculate t (in seconds).

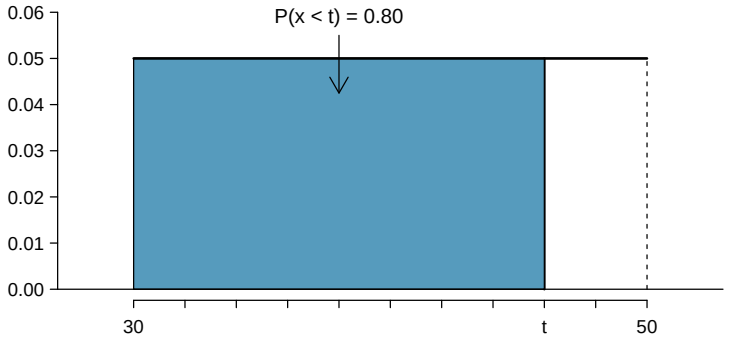


Figure 4.7: Plot of the continuous uniform distribution for Example (4.52) with the 80^{th} percentile marked as t.

To find t we can use calculus

$$\begin{aligned} P(30 \leq x \leq t) &= 0.80 \\ \int_{30}^{t} 0.05\,dx &= 0.80 \\ \left[0.05x\right]_{30}^{t} &= 0.80 \\ (0.05)(t) - (0.05)(30) &= 0.80 \\ t &= 46 \end{aligned}$$

We could also use the formula for the area of a rectangle

$$\begin{aligned} P(30 \leq x \leq t) &= 0.80 \\ (\text{base})(\text{height}) &= 0.80 \\ (t-30)(0.05) &= 0.80 \\ t &= 46 \end{aligned}$$

TIP: Identifying a process that follows a uniform distribution
When trying to identify a process that follows a continuous uniform distribution, look for the key words "equally likely" and a continuous finite interval in which the process occurs.

[20] $P(x \leq 45 | x > 40) = \frac{P(x \leq 45 \cap x > 40)}{P(x > 40)} = \frac{P(44 < x \leq 45)}{P(x > 40)} = \frac{(45-40)(0.05)}{(45-30)(0.05)} = 0.33$

The mean and variance of a uniform random variable are given below.

Mean and variance of a uniform random variable

Let $X \sim \mathrm{U}(a, b)$. The mean of X is

$$E(X) = \mu = \frac{a+b}{2} \tag{4.58}$$

and the variance of X is

$$V(X) = \sigma^2 = \frac{(b-a)^2}{12} \tag{4.59}$$

The result for the mean in (4.58) is intuitive since the distribution takes the form of a rectangle when plotted. The mean would therefore be the point which balances the rectangle perfectly and since the rectangle is symmetric the mean would be the midpoint of this rectangle. We can get the midpoint of a rectangle be taking the average of the two endpoints which is exactly what occurs in (4.58). We can also use (3.111) along with integral calculus to show how we obtain the man of this distribution.

Let X be a uniform random variable on the interval $[a, b]$. Then

$$E(X) = \mu = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx \tag{4.60}$$

$$= \int_a^b \frac{x}{b-a}\, dx \tag{4.61}$$

$$= \left[\frac{x^2}{2(b-a)}\right]_a^b \tag{4.62}$$

$$= \frac{b^2 - a^2}{2(b-a)} \tag{4.63}$$

$$= \frac{(b+a)(b-a)}{2(b-a)} \tag{4.64}$$

$$= \frac{a+b}{2} \tag{4.65}$$

Although (4.59) might appear strange with a denominator of 12 we can show how to obtain this result using (3.88). Again a knowledge of integral calculus is require for a full understanding of this derivation. Since we will be using (3.88) we start by calculating $E(X^2)$.

$$E(X^2) = \int_a^b \frac{x^2}{b-a}\,dx \tag{4.66}$$

$$= \left[\frac{x^3}{3(b-a)}\right]_a^b \tag{4.67}$$

$$= \frac{b^3 - a^3}{3(b-a)} \tag{4.68}$$

$$= \frac{(b-a)(b^2+ab+a^2)}{2(b-a)} \tag{4.69}$$

$$= \frac{b^2+ab+a^2}{3} \tag{4.70}$$

We can now substitute (4.70) into (3.88) and calculate the variance.

$$V(X) = \sigma^2 = E(X^2) - \mu^2 \tag{4.71}$$

$$= \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 \tag{4.72}$$

$$= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \tag{4.73}$$

$$= \frac{4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12} \tag{4.74}$$

$$= \frac{b^2-2ab+a^2}{12} \tag{4.75}$$

$$= \frac{(b-a)^2}{12} \tag{4.76}$$

Example 4.77 What are the mean and standard deviation for amount of time that the barista takes to make a cup of cappuccino

Using (4.58) the mean time taken is

$$\mu = \frac{50-30}{2}$$
$$= 30$$

The variance is

$$\sigma^2 = \frac{(50-30)^2}{12}$$
$$= 33.33$$

the standard deviation is therefore

$$\sigma = \sqrt{33.33}$$
$$= 5.77$$

4.2.2 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**,[21] shown in Figure 4.8. Variables such as SAT scores and heights of US adult males closely follow the normal distribution. There are many other real life processes and events that are modelled using the normal distribution.

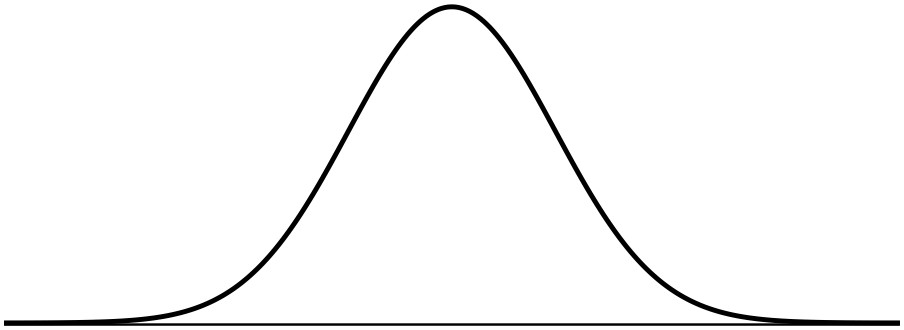


Figure 4.8: A normal curve.

Important properties of a normal distribution

1. Density function is completely described by μ and σ.
2. The normal distribution is symmetric about μ.
3. Mean = Median = Mode = μ

Aside from its use in statistics, the normal distribution is also used widely in many other areas (such as mathematics, econometrics, physics, biology, quantum mechanics, astronomy etc.

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.

4.2.2.1 Normal distribution model

The normal distribution model 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: the mean μ and

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

standard deviation σ. The density function of the normal distribution is explicitly given below.

Normal distribution

Let $X \sim N(\mu, \sigma^2)$. The density function of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < +\infty \tag{4.78}$$

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 4.9 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 4.10 shows these distributions on the same axis.

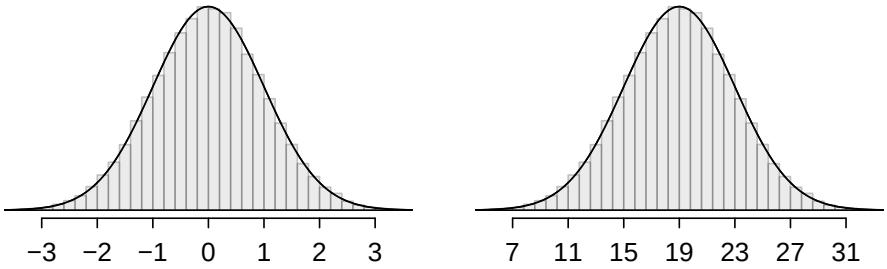


Figure 4.9: Both curves represent the normal distribution, however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

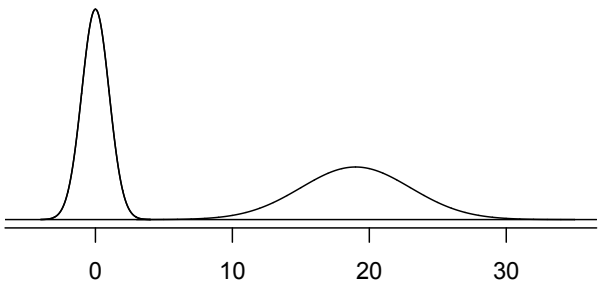


Figure 4.10: The normal models shown in Figure 4.9 but plotted together and on the same scale.

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 4.10 can be written as

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

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

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

⊙ **Exercise 4.79** Write down the short-hand for a normal distribution with (a) mean 5 and standard deviation 3, (b) mean -100 and standard deviation 10, and (c) mean 2 and standard deviation 9.[22]

Standard normal distribution

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1 (i.e. normal distribution where $\mu = 0$, $\sigma = 1$). We typically denote the standard normal distribution by Z.

$$Z \sim N(0, 1) \tag{4.80}$$

4.2.2.2 Standardizing with Z scores

In Section 4.2.2.1 where we discussed the normal distribution model we model a process as normal (or approximately normal) if the relative frequency distribution of the data appear to follow a symmetric bell curve. The density of this bell curve was also given by (4.78). The important point to note is that a process can be modelled as normal with any μ and any σ as long as its distribution is given by (4.78). Calculating probabilities for (4.78) is not as nice as it was for the continuous uniform distribution in Section 4.2.1. The integral of (4.78) is not trivial and requires knowledge beyond an introductory statistics class or even an introductory calculus class.

The way that we overcome this issue is by finding probabilities associated with the standard normal distribution and using these probabilities as a reference table of values. Whenever we encounter a problem in terms of a normal distribution any mean and standard deviation we can convert it to to an equivalent problem in terms of the standard normal distribution.

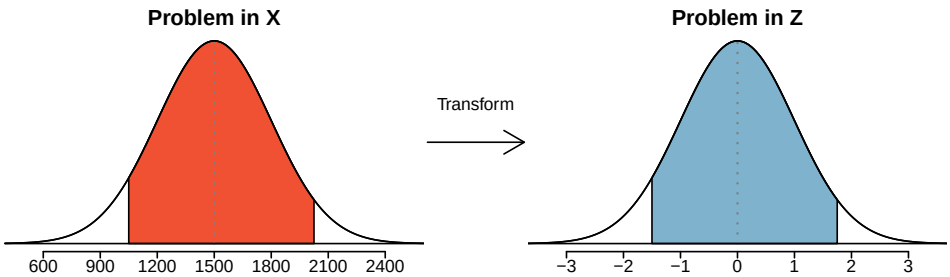


Figure 4.11: Transform a problem in terms of $X \sim N(\mu, \sigma^2)$ into an equivalent problem in terms of a $Z \sim N(0, 1)$.

We refer to our original problem as a problem in terms of random variable $X \sim N(\mu, \sigma^2)$ and the equivalent problem after the transformation as a problem in terms of random variable $Z \sim N(0, 1)$. The transformation used is given in Equation ((4.82)). Example (4.81) provides some motivation regarding how this transformation works.

● **Example 4.81** Table 4.12 on the facing page shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores

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

are both nearly normal. Suppose Ann scored 1800 on her SAT and Tom scored 24 on his ACT. Who performed better?

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT: $1500 + 300 = 1800$. Tom is 0.6 standard deviations above the mean on the ACT: $21 + 0.6 \times 5 = 24$. In Figure 4.13, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

	SAT	ACT
Mean	1500	21
SD	300	5

Table 4.12: Mean and standard deviation for the SAT and ACT.

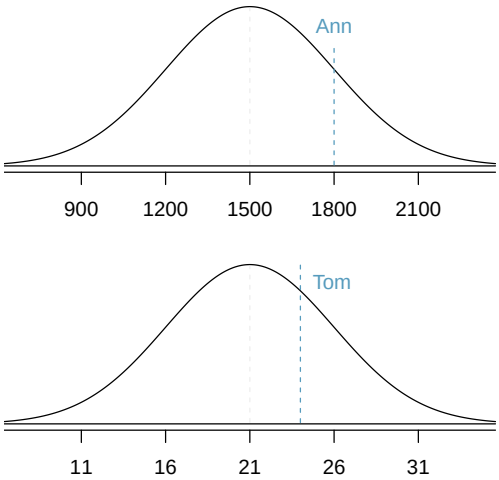


Figure 4.13: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

The Z score of an observation

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

$$Z = \frac{x - \mu}{\sigma} \tag{4.82}$$

Example (4.81) 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.

Z
Z score, the standardized observation

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

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

Example 4.83 Use Tom's ACT score, 24, along with the ACT mean and standard deviation to compute his Z score.

$$\begin{aligned} Z_{Tom} &= \frac{x_{Tom} - \mu_{ACT}}{\sigma_{ACT}} \\ &= \frac{24 - 21}{5} \\ &= 0.6 \end{aligned}$$

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

Exercise 4.84 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.[23]

Exercise 4.85 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.[24]

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

Exercise 4.86 Which of the observations in Exercise (4.85) is more unusual?[25]

Exercise 4.87 The variable `num_char` from the `email` data set describes the number of characters in nearly 4,000 emails. The distribution has mean 10,476 and standard deviation 14,383. Identify the Z scores for $\texttt{num\_char}_{36} = 13,788$ mm and $\texttt{num\_char}_{79} = 3,485$, which correspond to the 36^{th} and 76^{th} emails in the data set.[26]

Exercise 4.88 Which of the observations in Exercise (4.87) is more unusual?[27]

[23] (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.

[24] 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: $Z_2 = \frac{85.8-92.6}{3.6} = -1.89$.

[25] 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.

[26] $Z_{36} = \frac{13,788-10,476}{14,383} = 0.23$, $Z_{79} = \frac{3,485-10,476}{14,383} = -0.49$

[27] In Exercise (4.87), $Z_{36} = 0.23$ and $Z_{79} = -0.49$. Because the *absolute value* of Z_{79} is larger than Z_{36}, case 79 appears to have a more unusual number of email characters.

We can also use a transformation similar to (4.82) for the sample mean $\bar{x}$. We will learn more about the sampling distribution of the sample mean in Section 6.2.3 but we will give a gentle introduction to the topic since we work with another z score transformation. The basic idea is that that when a random sample of n measurements is selected from a population (following any distribution) with mean μ and standard deviation σ, the sample mean $\bar{x}$ becomes a random variable that follows a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. By this we mean that $\bar{x} \sim N(\mu, \sigma^2/n)$. We call $\sigma/\sqrt{n}$ the **standard error** of the mean.

The Z score of the sample average

The Z score of an average of n values is the number of standard deviations it falls above or below the mean. We compute the Z score for an average $\bar{x}$ that follows a distribution with mean μ and standard deviation σ using

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Example 4.89 Consider the SAT scores on in Table 4.12 on page 125. Suppose we took a random sample of 100 students who wrote the SAT. What is the Z score associated with a sample mean SAT score of 1870 of these 100 students?

From Table 4.12 we know that the mean of the SAT scores is $\mu = 1500$ and $\sigma = 300$ and we have a sample size of $n = 100$. Using Equation ((4.89))

$$\begin{aligned} Z &= \frac{1870 - 1800}{300/\sqrt{100}} \\ &= 2.33 \end{aligned}$$

Example 4.90 What would the sample mean of 64 students need to be in order for the average of these 64 students to be 2.5 standard deviations above the mean?

This means we are interested in a $\bar{x}$ when the Z score of 2.5. Using Equation ((4.89))

$$\begin{aligned} 2.5 &= \frac{\bar{x} - 1800}{300/\sqrt{64}} \\ \bar{x} &= 1893.75 \end{aligned}$$

4.2.2.3 Normal probability table

Example 4.91 Ann from Example (4.81) earned a score of 1800 on her SAT with a corresponding $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than Ann. We shade the area representing those individuals in Figure 4.14. The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in Figure 4.14: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

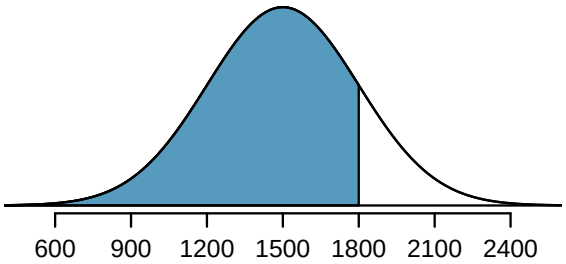


Figure 4.14: The normal model for SAT scores, shading the area of those individuals who scored below Ann.

We can use the normal model to find percentiles. A **normal probability table**, which lists Z scores and corresponding percentiles, can be used to identify a percentile based on the Z score (and vice versa). Statistical software can also be used.

A normal probability table is given in Appendix A on page 226 and abbreviated in Table 4.16. We use this table to identify the percentile corresponding to any particular Z score. For instance, the percentile of $Z = 0.43$ is shown in row 0.4 and column 0.03 in Table 4.16: 0.6664, or the 66.64^{th} percentile. Generally, we round Z to two decimals, identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation.

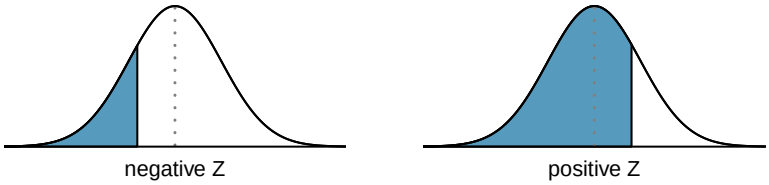


Figure 4.15: The area to the left of Z represents the percentile of the observation.

We can also find the Z score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z score for the 80^{th} percentile by combining the row and column Z values: 0.84.

⊙ **Exercise 4.92** Determine the proportion of SAT test takers who scored better than Ann on the SAT.[28]

4.2.2.4 Normal probability examples

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

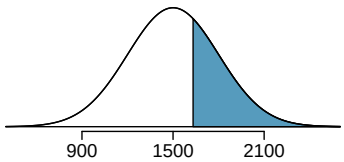
● **Example 4.93** Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1630 on her SATs?

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

	Second decimal place of Z									
Z	0.00	0.01	0.02	*0.03*	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	*0.6664*	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.16: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.43$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

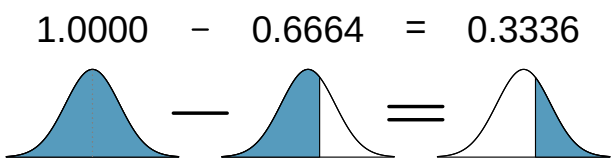
First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the chance she scores above 1630, so we shade this upper tail:



The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z score of the cutoff value. With $\mu = 1500$, $\sigma = 300$, and the cutoff value $x = 1630$, the Z score is computed as

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

We look up the percentile of $Z = 0.43$ in the normal probability table shown in Table 4.16 or in Appendix A on page 226, which yields 0.6664. However, the percentile describes those who had a Z score *lower* than 0.43. To find the area *above* $Z = 0.43$, we compute one minus the area of the lower tail:



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

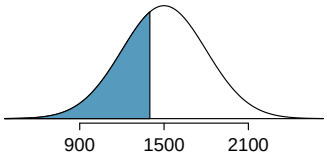
TIP: always draw a picture first, and find the Z score second
For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability.

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

⊙ **Exercise 4.94** If the probability of Shannon scoring at least 1630 is 0.3336, then what is the probability she scores less than 1630? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[29]

● **Example 4.95** Edward earned a 1400 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1400. These are the scores to the left of 1400.



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

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

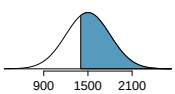
Using the normal probability table, identify the row of -0.3 and column of 0.03, which corresponds to the probability 0.3707. Edward is at the 37^{th} percentile.

⊙ **Exercise 4.96** Use the results of Example (4.95) to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[30]

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

[29] We found the probability in Example (4.93): 0.6664. A picture for this exercise is represented by the shaded area below "0.6664" in Example (4.93).

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



⊙ **Exercise 4.97** Stuart earned an SAT score of 2100. Draw a picture for each part. (a) What is his percentile? (b) What percent of SAT takers did better than Stuart?[31]

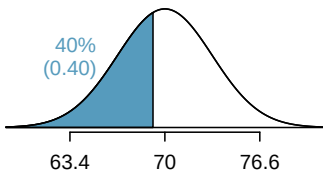
Based on a sample of 100 men,[32] the heights of male adults between the ages 20 and 62 in the US is nearly normal with mean 70.0" and standard deviation 3.3".

⊙ **Exercise 4.98** Mike is 5'7" and Jim is 6'4". (a) What is Mike's height percentile? (b) What is Jim's height percentile? Also draw one picture for each part.[33]

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

● **Example 4.99** Erik's height is at the 40^{th} percentile. How tall is he?

As always, first draw the picture.



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

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

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

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

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

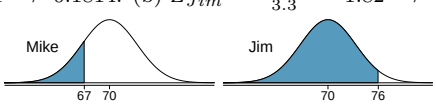
● **Example 4.100** What is the adult male height at the 82^{nd} percentile?

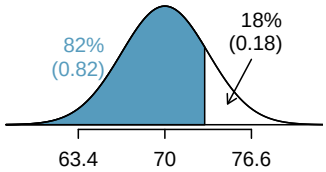
Again, we draw the figure first.

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

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

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





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

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

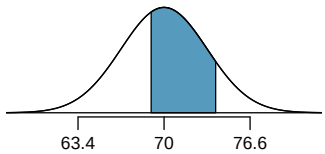
This yields 73.04 inches or about 6'1" as the height at the 82^{nd} percentile.

Exercise 4.101 (a) What is the 95^{th} percentile for SAT scores? (b) What is the 97.5^{th} percentile of the male heights? As always with normal probability problems, first draw a picture.[34]

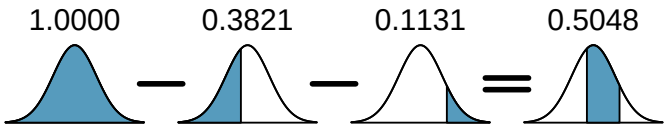
Exercise 4.102 (a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)? (b) What is the probability that a male adult is shorter than 5'9" (69 inches)?[35]

Example 4.103 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 Exercise (4.102), these areas are 0.3821 and 0.1131), then we can find the middle area:



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

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

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

⊙ **Exercise 4.104** What percent of SAT takers get between 1500 and 2000?[36]

⊙ **Exercise 4.105** What percent of adult males are between 5'5" and 5'7"?[37]

4.2.2.5 Empirical 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. This rule is called the **empirical rule** (it is also known as the "68-95-99.7" rule or the three-sigma rule).

Empirical rule

Let X be a random variable with a probability distribution that is approximately bell-shaped. Then

Approximately 66% of values lie within 1 standard deviation of the mean
Approximately 95% of values lie within 2 standard deviations of the mean
Approximately 99.7% of values lie within 3 standard deviations of the mean

In formal notation we mean that

$$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.66 \quad (4.106)$$
$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.95 \quad (4.107)$$
$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.997 \quad (4.108)$$

The empirical rule is illustrated in Figure 4.17.

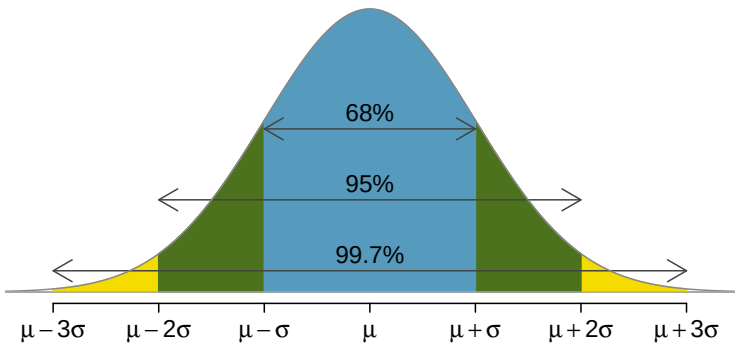


Figure 4.17: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

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

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

⊙ **Exercise 4.109** Use the Z table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[38]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-30,000. For 5 and 6 standard deviations, it is about 1-in-3.5 million and 1-in-1 billion, respectively.

⊙ **Exercise 4.110** SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$. (a) About what percent of test takers score 900 to 2100? (b) What percent score between 1500 and 2100?[39]

4.2.2.6 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 4.111** Approximately 20% 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 59 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 20%, what is the probability of observing 59 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.

The question posed is equivalent to asking, what is the probability of observing $k = 0$, 1, ..., 58, or 59 smokers in a sample of $n = 400$ when $p = 0.20$? We can compute these 60 different probabilities and add them together to find the answer:

$$\begin{aligned} &P(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 59) \\ &\quad = P(k = 0) + P(k = 1) + \cdots + P(k = 59) \\ &\quad = 0.0041 \end{aligned}$$

If the true proportion of smokers in the community is $p = 0.20$, then the probability of observing 59 or fewer smokers in a sample of $n = 400$ is less than 0.0041.

The computations in Example (4.111) 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.

[38] First draw the pictures. To find the area between $Z = -1$ and $Z = 1$, use the normal probability table to determine the areas below $Z = -1$ and above $Z = 1$. Next verify the area between $Z = -1$ and $Z = 1$ is about 0.68. Repeat this for $Z = -2$ to $Z = 2$ and also for $Z = -3$ to $Z = 3$.

[39] (a) 900 and 2100 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 900 and 2100. (b) Since the normal model is symmetric, then half of the test takers from part (a) ($\frac{95\%}{2} = 47.5\%$ of all test takers) will score 900 to 1500 while 47.5% score between 1500 and 2100.

⊙ **Exercise 4.112** Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 4.18 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?[40]

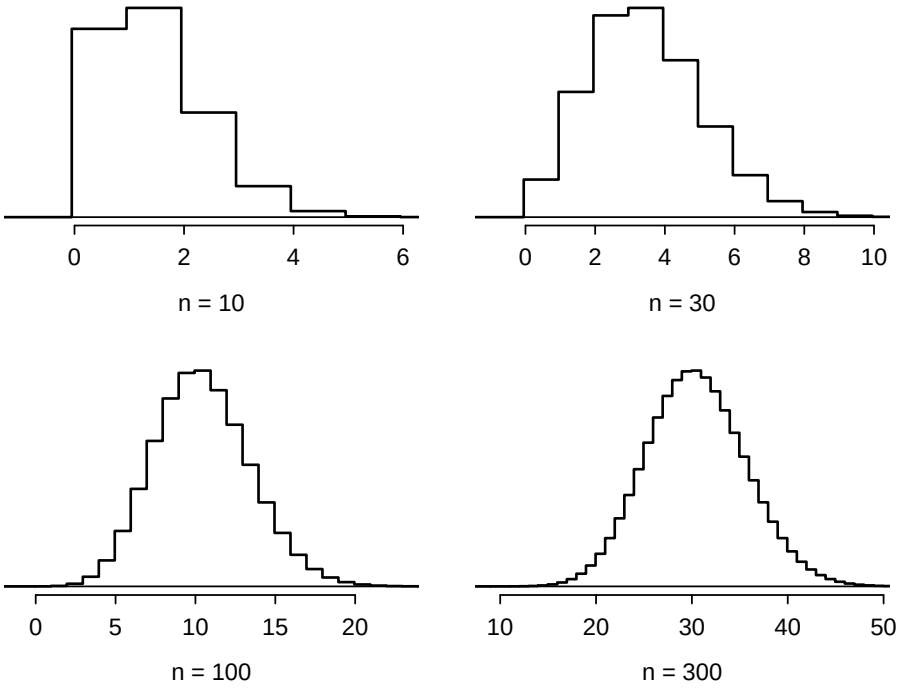


Figure 4.18: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

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:

$$E(X) = \mu = np \tag{4.113}$$

and

$$V(X) = \sigma^2 = np(1-p) \tag{4.114}$$

[40] The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in last hollow histogram

From (4.114) we see that the standard deviation of the approximated normal distribution is clearly $\sigma = \sqrt{np(1-p)}$ 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 (4.111).

Example 4.115 How can we use the normal approximation to estimate the probability of observing 59 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.20$?

Showing that the binomial model is reasonable was a suggested exercise in Example (4.111). We also verify that both np and $n(1-p)$ are at least 10:

$$np = 400 \times 0.20 = 80 \qquad n(1-p) = 400 \times 0.8 = 320$$

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:

$$\mu = np = 80 \qquad \sigma = \sqrt{np(1-p)} = 8$$

We want to find the probability of observing fewer than 59 smokers using this model.

Exercise 4.116 Use the normal model $N(\mu = 80, \sigma = 8)$ to estimate the probability of observing fewer than 59 smokers. Your answer should be approximately equal to the solution of Example (4.111): 0.0041.[41]

4.2.2.7 The normal approximation breaks down on small intervals

Caution: The normal approximation may fail on small intervals
The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 69, 70, or 71 smokers in 400 when $p = 0.20$. With such a large sample, we might be tempted to apply the normal approximation and use the range 69 to 71. However, we would find that the binomial solution and the normal approximation notably differ:

Binomial: 0.0703 Normal: 0.0476

We can identify the cause of this discrepancy using Figure 4.19, 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.

[41] Compute the Z score first: $Z = \frac{59-80}{8} = -2.63$. The corresponding left tail area is 0.0043.

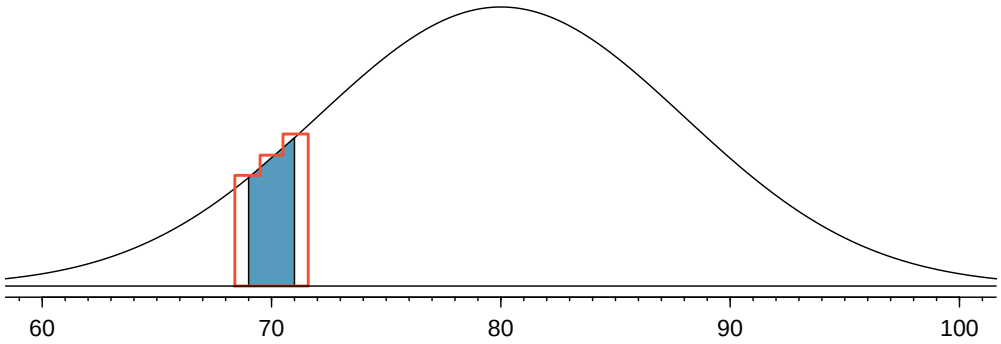


Figure 4.19: A normal curve with the area between 69 and 71 shaded. The outlined area represents the exact binomial probability.

TIP: Improving the accuracy of the normal approximation to the binomial distribution
The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values 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. While it is possible to apply it when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

4.2.2.8 Normal probability plots

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males. While using a normal model can be extremely convenient and helpful, it is important to remember normality is always an approximation. Testing the appropriateness of the normal assumption is a key step in many data analyses.

Example (4.99) suggests the distribution of heights of US males is well approximated by the normal model. We are interested in proceeding under the assumption that the data are normally distributed, but first we must check to see if this is reasonable.

There are two visual methods for checking the assumption of normality, which can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 4.20. The sample mean $\bar{x}$ and standard deviation s are used as the parameters of the best fitting normal curve. The closer this curve fits the histogram, the more reasonable the normal model assumption. Another more common method is examining a **normal probability plot**.[42], shown in the right panel of Figure 4.20. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model. We outline the construction of the normal probability plot in Section 4.2.2.9

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

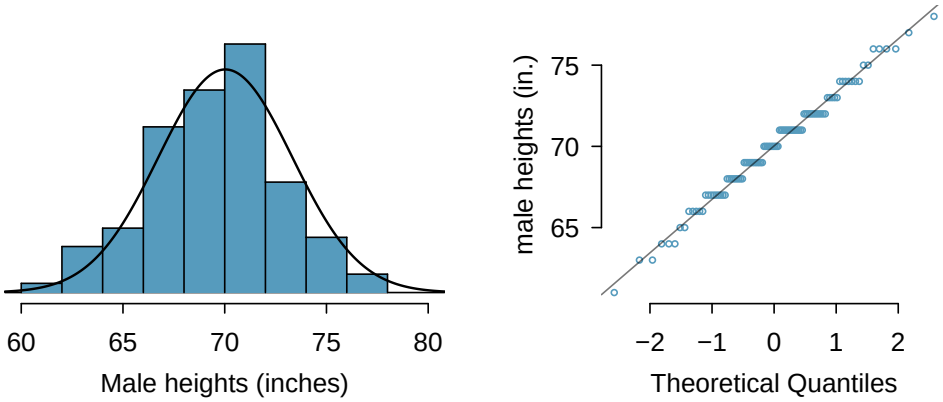


Figure 4.20: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

Example 4.117 Three data sets of 40, 100, and 400 samples were simulated from a normal distribution, and the histograms and normal probability plots of the data sets are shown in Figure 4.21. These will provide a benchmark for what to look for in plots of real data.

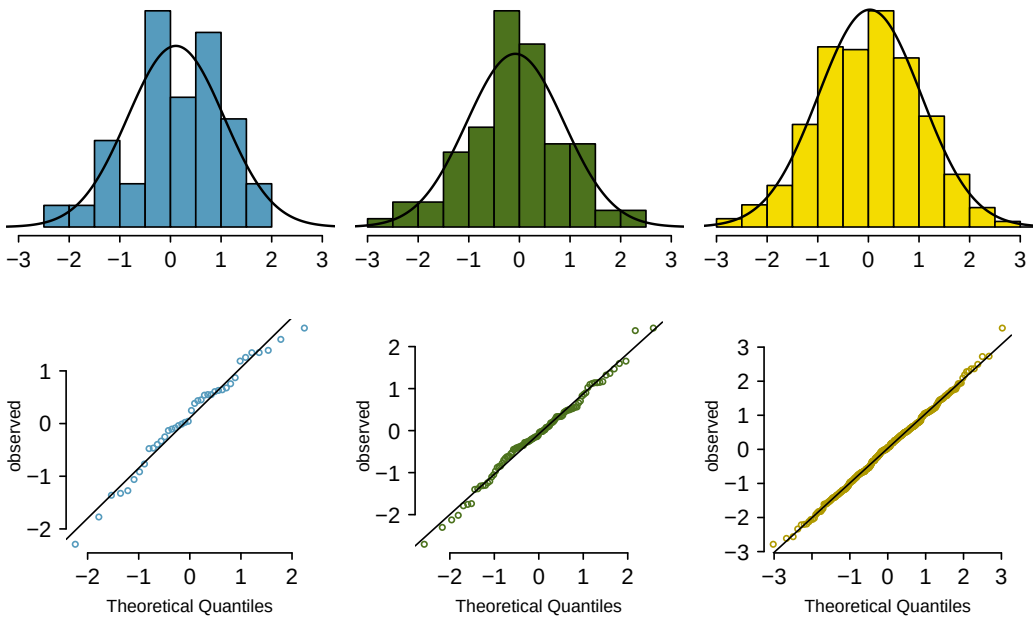


Figure 4.21: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

The left panels show the histogram (top) and normal probability plot (bottom) for the simulated data set with 40 observations. The data set is too small to really see

clear structure in the histogram. The normal probability plot also reflects this, where there are some deviations from the line. However, these deviations are not strong.

The middle panels show diagnostic plots for the data set with 100 simulated observations. The histogram shows more normality and the normal probability plot shows a better fit. While there is one observation that deviates noticeably from the line, it is not particularly extreme.

The data set with 400 observations has a histogram that greatly resembles the normal distribution, while the normal probability plot is nearly a perfect straight line. Again in the normal probability plot there is one observation (the largest) that deviates slightly from the line. If that observation had deviated 3 times further from the line, it would be of much greater concern in a real data set. Apparent outliers can occur in normally distributed data but they are rare.

Notice the histograms look more normal as the sample size increases, and the normal probability plot becomes straighter and more stable.

Example 4.118 Are NBA player heights normally distributed? Consider all 435 NBA players from the 2008-9 season presented in Figure 4.22.[43]

We first create a histogram and normal probability plot of the NBA player heights. The histogram in the left panel is slightly left skewed, which contrasts with the symmetric normal distribution. The points in the normal probability plot do not appear to closely follow a straight line but show what appears to be a "wave". We can compare these characteristics to the sample of 400 normally distributed observations in Example (4.117) and see that they represent much stronger deviations from the normal model. NBA player heights do not appear to come from a normal distribution.

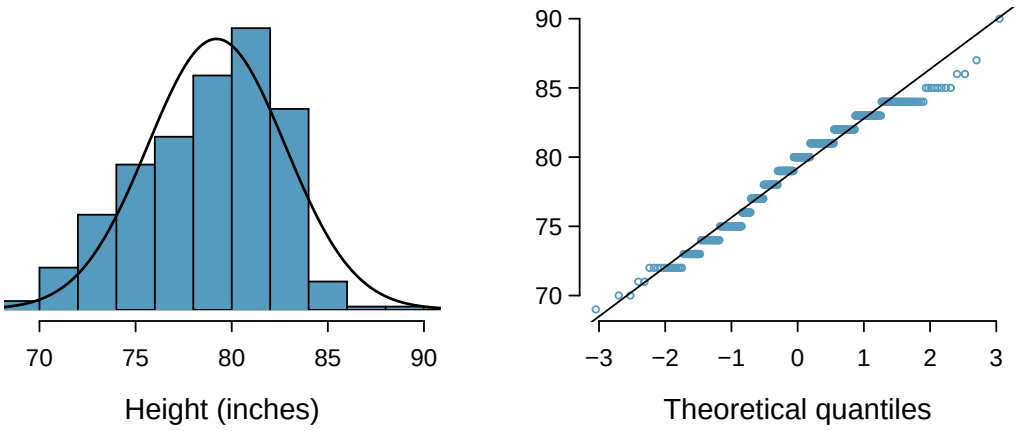


Figure 4.22: Histogram and normal probability plot for the NBA heights from the 2008-9 season.

Example 4.119 Can we approximate poker winnings by a normal distribution? We consider the poker winnings of an individual over 50 days. A histogram and normal probability plot of these data are shown in Figure 4.23.

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

The data are very strongly right skewed in the histogram, which corresponds to the very strong deviations on the upper right component of the normal probability plot. If we compare these results to the sample of 40 normal observations in Example (4.117), it is apparent that these data show very strong deviations from the normal model.

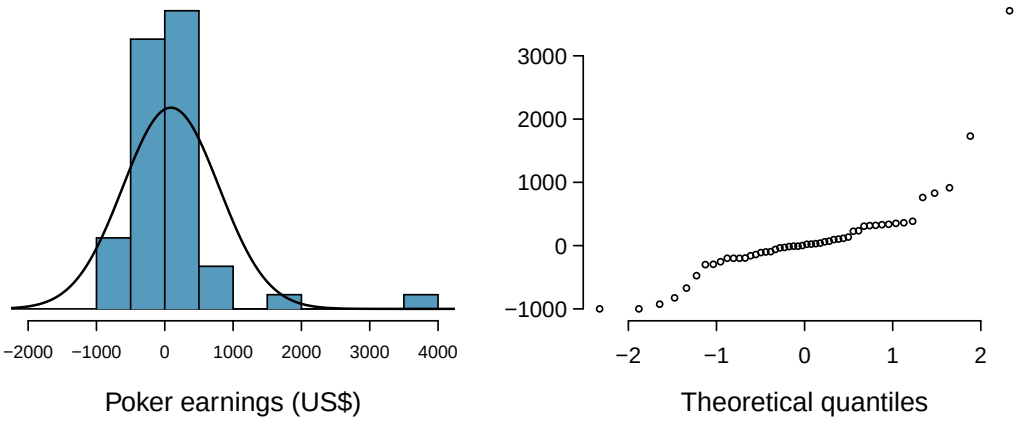


Figure 4.23: A histogram of poker data with the best fitting normal plot and a normal probability plot.

⊙ **Exercise 4.120** If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[44]

⊙ **Exercise 4.121** Determine which data sets represented in Figure 4.24 plausibly come from a nearly normal distribution. Are you confident in all of your conclusions? There are 100 (top left), 50 (top right), 500 (bottom left), and 15 points (bottom right) in the four plots.[45]

[44] 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 7.5.) To determine this probability, look up -2.58 and 2.58 in the normal probability table (0.0049 and 0.9951). Thus, there is a $0.9951 - 0.0049 \approx 0.99$ probability that the unobserved random variable X will be within 2.58 standard deviations of μ.

[45] Answers may vary a little. The top-left plot shows some deviations in the smallest values in the data set; specifically, the left tail of the data set has some outliers we should be wary of. The top-right and bottom-left plots do not show any obvious or extreme deviations from the lines for their respective sample sizes, so a normal model would be reasonable for these data sets. The bottom-right plot has a consistent curvature that suggests it is not from the normal distribution. If we examine just the vertical coordinates of these observations, we see that there is a lot of data between -20 and 0, and then about five observations scattered between 0 and 70. This describes a distribution that has a strong right skew.

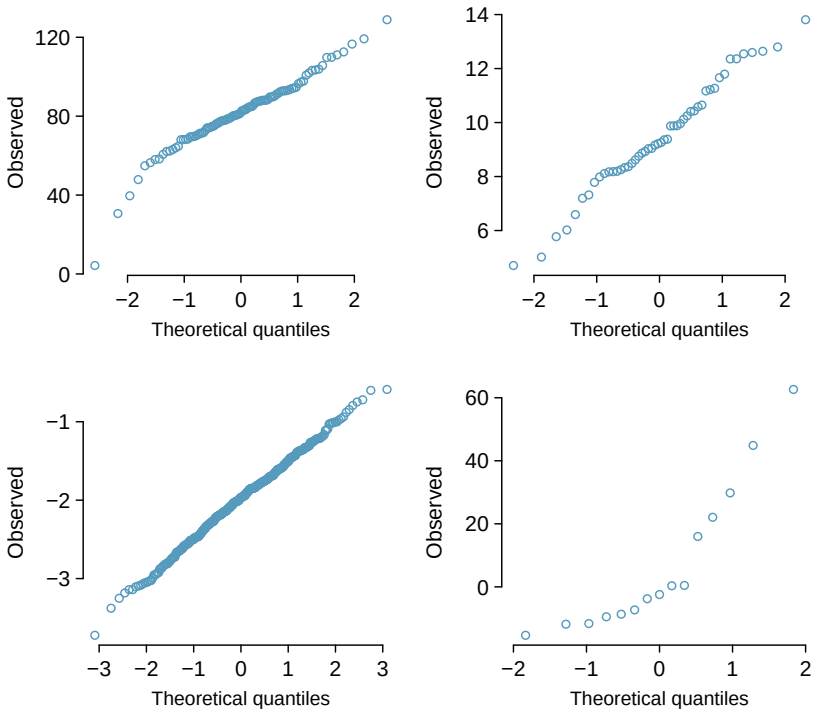


Figure 4.24: Four normal probability plots for Exercise (4.121).

⊙ **Exercise 4.122** Figure 4.25 shows normal probability plots for two distributions that are skewed. One distribution is skewed to the low end (left skewed) and the other to the high end (right skewed). Which is which?[46]

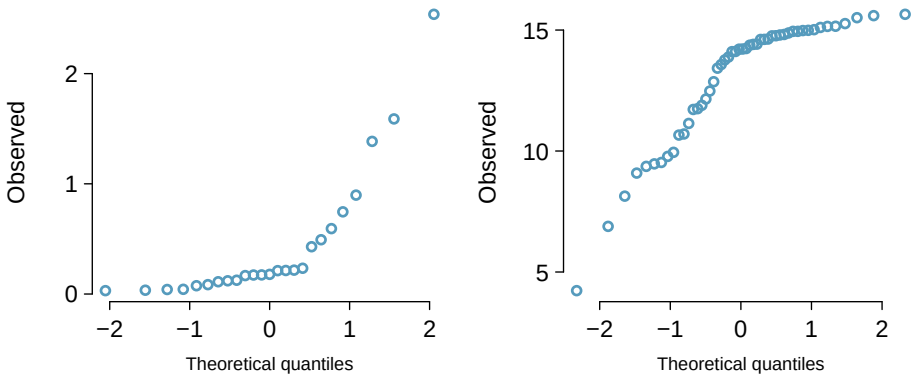


Figure 4.25: Normal probability plots for Exercise (4.122).

[46]Examine where the points fall along the vertical axis. In the first plot, most points are near the low end with fewer observations scattered along the high end; this describes a distribution that is skewed to the high end. The second plot shows the opposite features, and this distribution is skewed to the low end.

4.2.2.9 Constructing a normal probability plot (special topic)

We construct a normal probability plot for the heights of a sample of 100 men as follows:

(1) Order the observations.

(2) Determine the percentile of each observation in the ordered data set.

(3) Identify the Z score corresponding to each percentile.

(4) Create a scatterplot of the observations (vertical) against the Z scores (horizontal).

If the observations are normally distributed, then their Z scores will approximately correspond to their percentiles and thus to the z_i in Table 4.26.

Observation i	1	2	3	$\cdots$	100
x_i	61	63	63	$\cdots$	78
Percentile	0.99%	1.98%	2.97%	$\cdots$	99.01%
z_i	-2.33	-2.06	-1.89	$\cdots$	2.33

Table 4.26: Construction details for a normal probability plot of 100 men's heights. The first observation is assumed to be at the 0.99^{th} percentile, and the z_i corresponding to a lower tail of 0.0099 is -2.33. To create the plot based on this table, plot each pair of points, (z_i, x_i).

Caution: z_i correspond to percentiles
The z_i in Table 4.26 are *not* the Z scores of the observations but only correspond to the percentiles of the observations.

Because of the complexity of these calculations, normal probability plots are generally created using statistical software.

4.2.3 t distribution

The Student's t distribution, or t distribution for short, is a continuous family of probability distributions that are very useful in statistical inference. It is typically used when the standard deviation is not known.

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.

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.

The degrees of freedom refer to the number of free values that we can vary.

TIP: Note on degrees of freedom of a t distribution
Strictly speaking the the degrees of freedom can be any positive real value and do have be positive integers. However for this introductory text all our exercises and examples will involve the degrees of freedom as positive integers.

This definition will become more clear in Section **??**. We provide the density of the t distribution for completeness.

t distribution

Let X be a random variable that follows a t distribution with r degrees of freedom. In standard notation: $X \sim t_r$. The density function of X is

$$f_r(x) = \frac{\Gamma\left[\frac{1}{2}(r+1)\right]}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)\left(1+\frac{x^2}{r}\right)^{(r+1)/2}} \qquad -\infty < x < +\infty \tag{4.123}$$

where $\Gamma(x)$ is the gamma function defined as

$$\Gamma(x) = \int_0^\infty t^{x-1}e^t\,dt \tag{4.124}$$

Although the density function may appear complicated the shape of the t distribution is familiar. The t distribution resembles the normal distribution, except with thicker tails in general. The shape of the t distribution varies with the degrees of freedom. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution. Just like the normal distribution, all t distributions are symmetric.

A t distribution is shown as a solid line in Figure 4.27. The red dotted line is the standard normal distribution. Notice the familiar a bell shape, but also notice that its tails are thicker than those of the normal model. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[47] We will learn that these extra thick tails will be useful in the section on colourredinference.

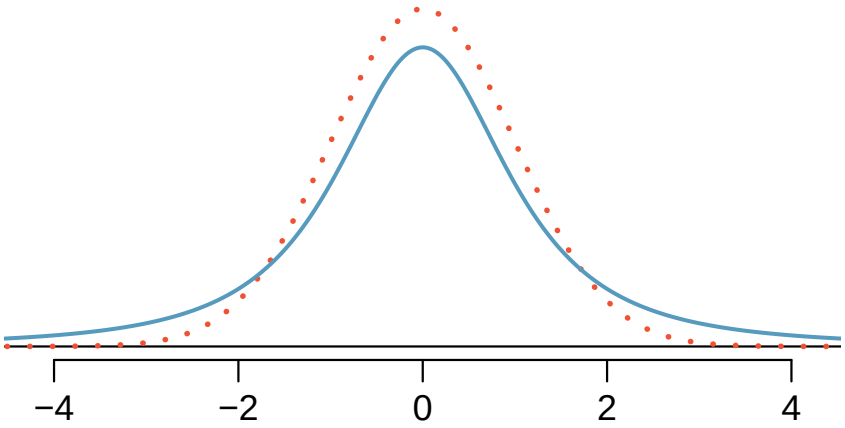


Figure 4.27: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

[47] 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.

Several t distributions are shown in Figure 4.28.

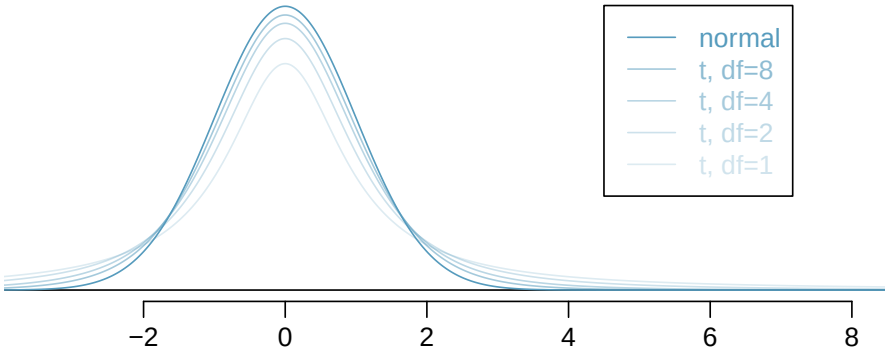


Figure 4.28: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

When the degrees of freedom are about 30 or more, the t distribution is almost indistinguishable from the normal distribution. In Section **??**, we relate degrees of freedom to sample size.

4.2.3.1 t table

We will find it very useful to become familiar with the t distribution, because it plays a very similar role to the normal distribution during inference for small samples of numerical data. We use a **t table**, partially shown in Table 4.29, in place of the normal probability table for small sample numerical data. A larger table is presented in Appendix A.3 on page 232.

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

Table 4.29: An abbreviated look at the t table. Each row represents a different t distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

Each row in the t table represents a t distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t distribution with $df = 18$, we examine row 18, which is **highlighted** in Table 4.29. If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where *one tail* is 0.100. In standard notation we are interested in $t_{(0.10,\,18)}$. So our cutoff vale from the t table is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33.

In the example above we had $t_{(0.10,\,18)} = 1.33$. Notice that the standard notation of the value taken from t distribution had two subscripts. The first subscript indicates the area in the tail and we usually refer to this as $\frac{\alpha}{2}$. The second subscript were the degrees of freedom. It is a good idea to get familiar with this notation as we will be using it more in subsequent chapters, particularly in Section 7 and 8.

Example 4.125 What proportion of the t distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 4.30 and shade the area below -2.10. To find this area, we identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

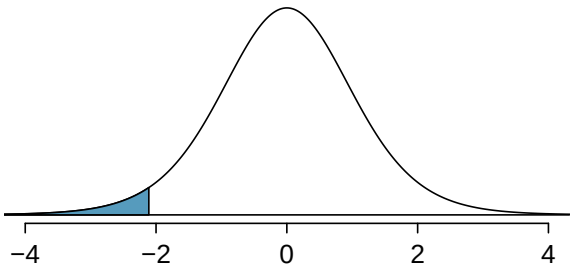


Figure 4.30: The t distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

Example 4.126 A t distribution with 20 degrees of freedom is shown in the left panel of Figure 4.31. Estimate the proportion of the distribution falling above 1.65.

We identify the row in the t table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

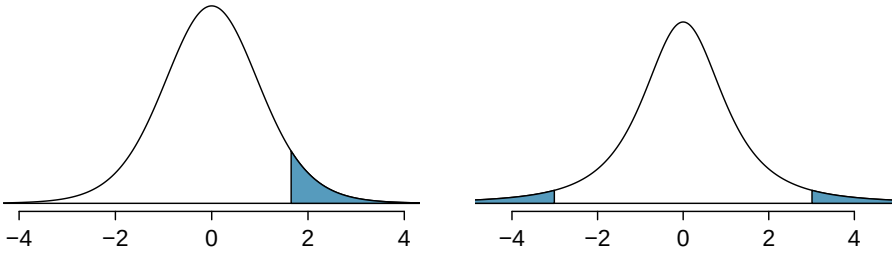


Figure 4.31: Left: The t distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

Example 4.127 A t distribution with 2 degrees of freedom is shown in the right panel of Figure 4.31. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: $df = 2$. Next, find the columns that capture 3; because $2.92 < 3 < 4.30$, we use the second and third columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

Exercise 4.128 What proportion of the t distribution with 19 degrees of freedom falls above -1.79 units?[48]

[48] 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.

Chapter 5

Basic foundations of Inference

Statistical inference is concerned primarily with understanding the quality of parameter estimates. For example, a classic inferential question is, "How sure are we that the estimated mean, $\bar{x}$, is near the true population mean, μ?" These are the questions that are answered in sections REFERENCE but at this point we set the stage for understanding the procedures behind these sections. A good familiarity with this chapter will make the rest of this book, and indeed the rest of statistics, seem much more familiar. Sections 5.1, 5.2 and 6.1 all come together nicely in the example presented in case study 5.4. While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics.

5.1 Sampling distributions

In this section we discuss some background material that we use for Chapters 6, **??** and **??**. The material in this section is not a rigorous exploration into sampling theory but it should be sufficient enough for a good foundation to understand the material in Chapters 6, **??** and **??**.

We start with defining a point estimate. A point estimate is a single value that can be regarded as the best guess of a parameter.

Point Estimate
A **point estimate** is a a single value (i.e. a single point on the real number line) that estimates the value of a parameter

A point estimate is obtained by choosing a suitable statistic that is calculated from sample data. The selected statistic is called the **point estimator** of the parameter. For example the sample mean $\bar{x}$ is a point estimate of the population mean μ, and the sample standard deviation s^2 is a point estimate of σ^2 which is the population variance [1]. Methods of finding point estimates such as maximum likelihood estimation and the method of moments can be learnt in a course on mathematical statistics.

The next important term to define is **the sampling distribution**. A sampling distribution is the distribution of a statistic.

[1] note that there are other point estimators available σ^2 such as $\frac{(n-1)}{n}s^2$

Sampling distribution
The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from such a distribution.

Suppose we are interested in a certain population parameter θ. and our best guess of θ is point estimator $\hat{\theta}$. We are now interested in creating a sampling distribution for $\hat{\theta}$. A sample of data of size n is drawn from the population and a single instance of the estimator $\hat{\theta}_1$ is calculated from this sample data. Another **independent sample** which is also size n is drawn and another instance of the estimator $\hat{\theta}_1$ is calculated from this different sample data. This process is repeated many times and we obtain a collection estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ where m is very large; such as the order of several thousands[2]. We can now create a relative frequency distribution for the large collection of samples obtained. As m approaches ∞ (i.e. we take an infinite number of samples and calculate a statistic for each one of these samples) the relative frequency distribution created for $\hat{\theta}$ is the sampling distribution of $\hat{\theta}$. Figure 5.1 below gives a summary of the manner in which a sampling distribution is created.

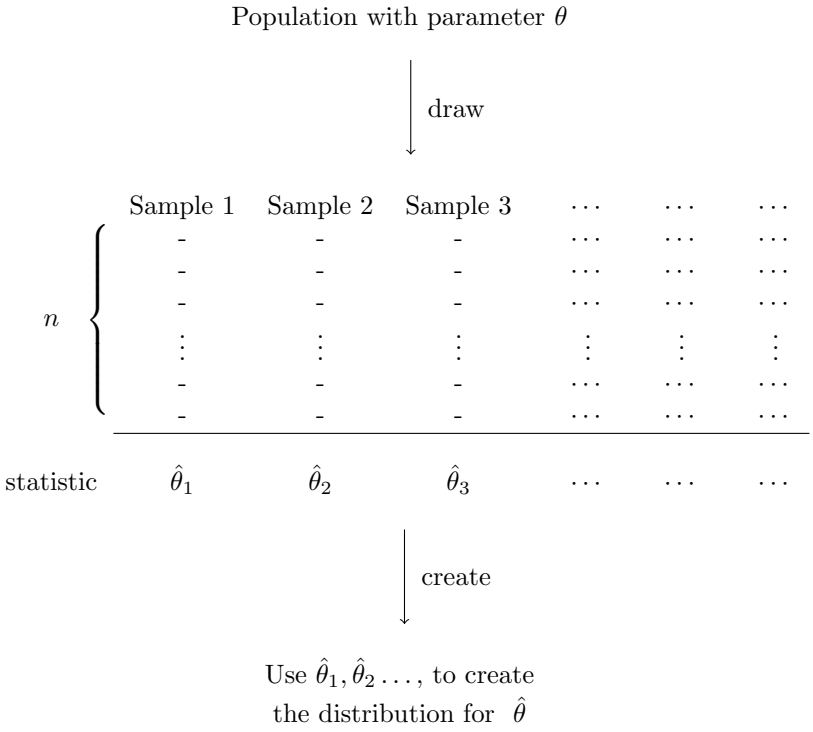


Figure 5.1: Schematic summary on creating a sampling distribution for an estimator $\hat{\theta}$.

We use this framework to create a sampling distribution for the sample mean $\bar{x}$. which is of particular interest to us. In the next section we discuss a very important theorem

[2]It is important to note that samples drawn are independent of one another. Whether samples are drawn with or without replacement usually does not matter since a population is typically very large.

related to the sample distributions which is the Central Limit Theorem. The central limit theorem is used as a foundation for a lot of inference techniques.

5.2 Central Limit Theorem

The normal model for the sample mean tends to be very good when the sample consists of at least a sufficiently large number[3] of independent observations and the population data are not strongly skewed. The Central Limit Theorem provides the theory that allows us to make this assumption.

Central Limit Theorem (CLT)
Consider a random sample of size n drawn from a population having mean μ and standard deviation σ. (The population may follow any underlying distribution, however we know its mean is μ and its standard deviation is σ). For sufficiently large n the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed with a mean of $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

$$\bar{x} \overset{CLT}{\sim} N(\mu, \sigma/n)$$

Stated informally, the central limit theorem states that the distribution of $\bar{x}$ is approximately normal with mean μ and variance σ/n where n is the size of the sample drawn. The population from which the samples are drawn can follow any distribution, however after going through the procedure outlined in Figure 5.1, the distribution of $\bar{x}$ will be normal. Note that the approximation can be poor if n is small, but it improves with larger sample sizes[4].

The definition above is one way of stating the classical central limit theorem. We are not being extremely formal since the classical central limit theorem can be expressed in terms of $\bar{x}$ converging in distribution to $N(\mu, \sigma/\sqrt{n}$, however this is something that is more suited for a course in statistical inference of mathematical statistics. There are other ways of defining the central limit theorem and there are special forms of the central limit theorem such as the Lindeberg-Levy CLT, however the definition above is the most appropriate one for this course.

We will investigate three cases to see (roughly) when the approximation is reasonable. We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *log-normal* distribution. These distributions are shown in the top panels of Figure 5.2. The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

[3] The definition of "sufficiently large" is subjective. However we shall consider a sample size of $n \geq 30$ to be large.

[4] n should not be confused with m mentioned in 5.1

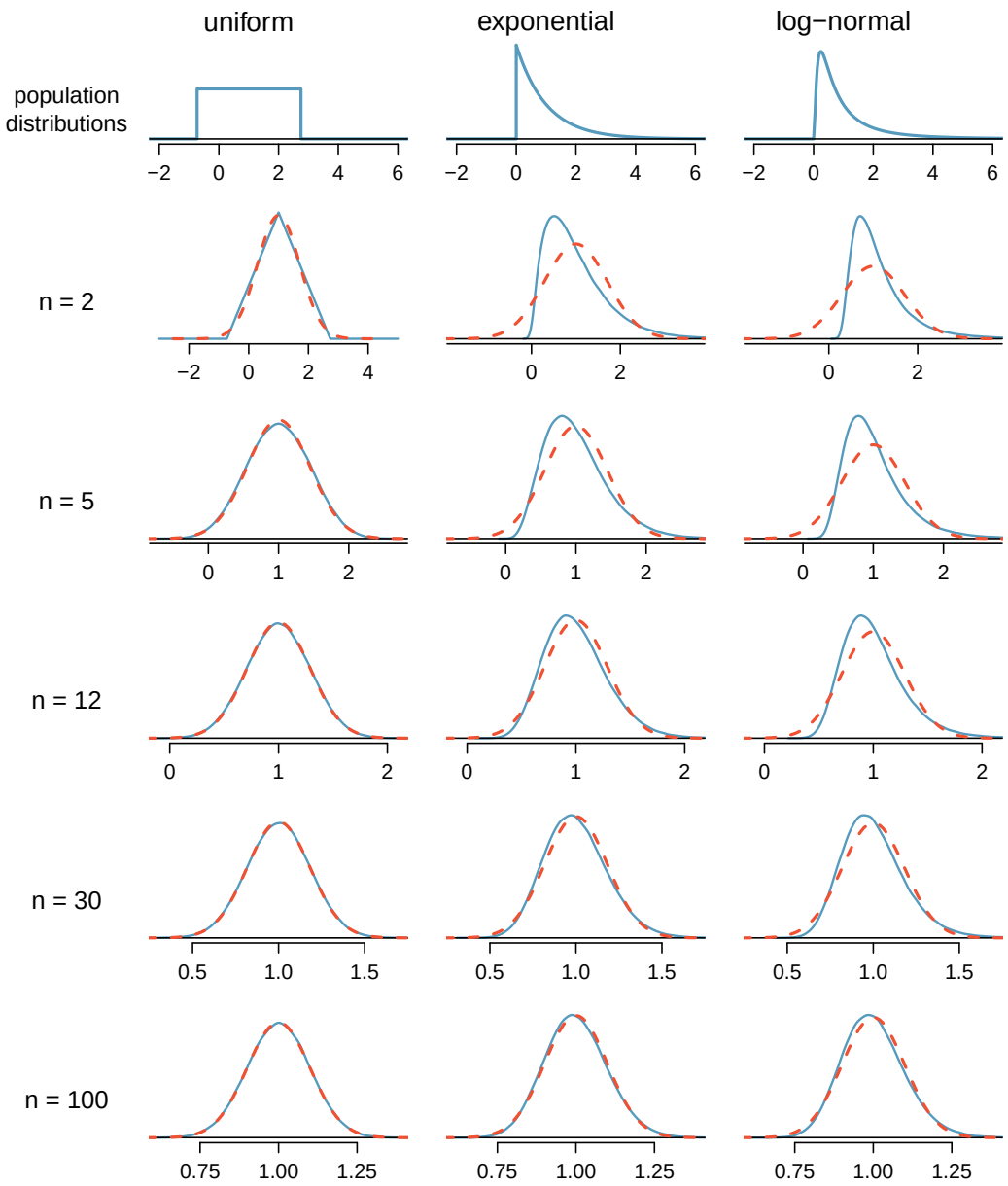


Figure 5.2: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

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

⊙ **Exercise 5.1** Examine the distributions in each row of Figure 5.2. What do you notice about the normal approximation for each sampling distribution as the sample size becomes larger?[5]

● **Example 5.2** Would the normal approximation be good in all applications where the sample size is at least 30?

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

TIP: With larger n, the sampling distribution of $\bar{x}$ becomes more normal
As the sample size increases, the normal model for $\bar{x}$ becomes more reasonable.

● **Example 5.3** Figure 5.3 shows a histogram of 50 observations.

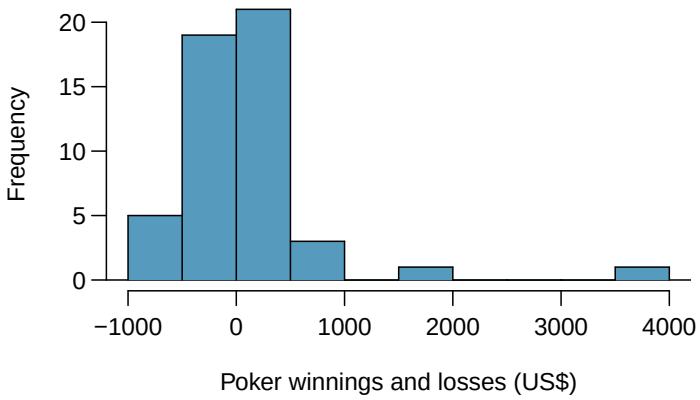


Figure 5.3: Sample distribution of poker winnings. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

These represent winnings and losses from 50 consecutive days of a professional poker player. Can the normal approximation be applied to the sample mean, 90.69?

We should consider each of the required conditions.

(1) These are referred to as **time series data**, because the data arrived in a particular sequence. If the player wins on one day, it may influence how she plays the next. To make the assumption of independence we should perform careful checks on such data. While the supporting analysis is not shown, no evidence was found to indicate the observations are not independent.

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

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

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

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

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

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

5.3 Variability in point estimates

The problem with point estimates is that they are usually not exactly equal to the value of the population parameter they are estimating. Due to the nature of randomness the chances are that the point estimate taken from a single sample will not be exact equal to the population parameter of interest. It is important to note that the value of the population parameter stays fixed although the value of point estimates drawn from samples can vary.

Parameter values do not vary
We assume that the value of a population parameter remains fixed.

For example, if we are interested in the mean height of all students at a university with a population of 20,000 students. We assume that the true mean height of the 20,000 students is some fixed value that does not change. Suppose we take a random sample of 30 students and measure the heights of these students. There is a good chance that the mean of the 30 students will not be exact equal to the mean of all 20,000 students. As we increase our sample size then our estimate will become a better approximation of the true parameter. Going back to our example of measuring student height, suppose we calculated a sample mean with the 100 students instead of 30. The sample mean based on 100 students is a better approximation of the true population mean.

The way that we can quantify the amount of uncertainty of an estimator is with the **standard error** of the estimator.

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

The standard error is the standard deviation of the sampling distribution of the statistic of interest. The calculation of the standard error depends on the situation. For instance lets go back to the central limit theorem in Section 5.2. We saw that the sampling distribution for the sample mean $\bar{x}$ is a normal distribution with mean μ and variance $\sigma^2/\sqrt{n}$. Therefore the standard deviation of the sampling distribution of $\bar{x}$ is $\sigma/\sqrt{n}$ which means that the standard error of the mean is $\sigma/\sqrt{n}$.

We will learn the standard error of different point estimates in Chapters REFERENCE of this text.

5.4 Case study: Cherry blosson 10 mile run

We will illustrate the central limit theorem using a case study with with real data. Throughout the next few sections we consider a data set called `run10`, which represents all 16,924 runners who finished the 2012 Cherry Blossom 10 mile run in Washington, DC.[6] Part of this data set is shown in Table 6.1, and the variables are described in Table 6.2.

ID	time	age	gender	state
1	92.25	38.00	M	MD
2	106.35	33.00	M	DC
3	89.33	55.00	F	VA
4	113.50	24.00	F	VA
⋮	⋮	⋮	⋮	⋮
16923	122.87	37.00	F	VA
16924	93.30	27.00	F	DC

Table 5.4: Six observations from the `run10` data set.

variable	description
`time`	Ten mile run time, in minutes
`age`	Age, in years
`gender`	Gender (`M` for male, `F` for female)
`state`	Home state (or country if not from the US)

Table 5.5: Variables and their descriptions for the `run10` data set.

These data are special because they include the results for the entire population of runners who finished the 2012 Cherry Blossom Run. We took a simple random sample of 100 observations from this population, which is represented in Table 6.3. We will use this sample, which we refer to as the `run10Samp` data set, to draw conclusions about the entire population. This is the practice of statistical inference in the broadest sense. Two histograms summarizing the time and age variables in the `run10Samp` data set are shown in Figure 6.4.

[6] http://www.cherryblossom.org

Sample observation	ID	time	age	gender	state
1	1983	88.31	59	M	MD
2	8192	100.67	32	M	VA
3	11020	109.52	33	F	VA
⋮	⋮	⋮	⋮	⋮	⋮
100	1287	89.49	26	M	DC

Table 5.6: Four observations for the `run10Samp` data set, which represents a simple random sample of 100 runners from the 2012 Cherry Blossom Run.

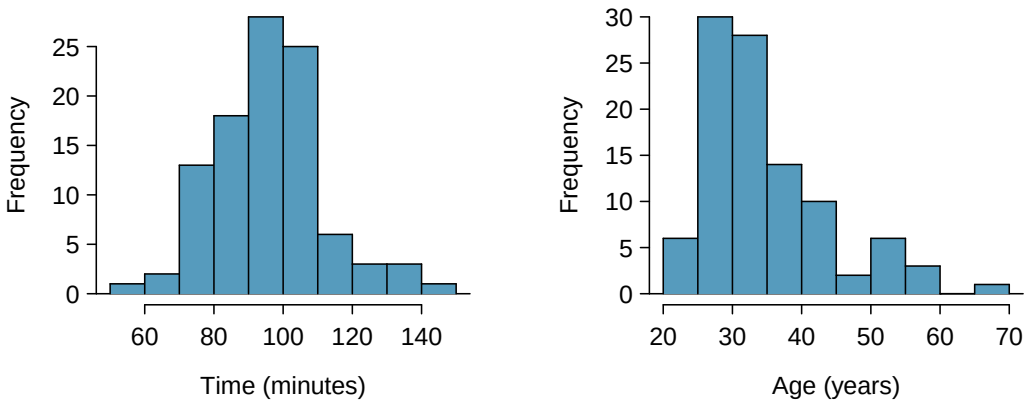


Figure 5.7: Histograms of `time` and `age` for the sample Cherry Blossom Run data. The average time is in the mid-90s, and the average age is in the mid-30s. The age distribution is moderately skewed to the right.

We would like to estimate two features of the Cherry Blossom runners using the sample.

(1) How long does it take a runner, on average, to complete the 10 miles?

(2) What is the average age of the runners?

These questions may be informative for planning the Cherry Blossom Run in future years.[7] We will use $x_1, ..., x_{100}$ to represent the 10 mile time for each runner in our sample, and $y_1, ..., y_{100}$ will represent the age of each of these participants.

5.4.1 Calculations of the sample mean

We want to estimate the **population mean** based on the sample. The most intuitive way to go about doing this is to simply take the **sample mean**. That is, to estimate the average 10 mile run time of all participants, take the average time for the sample:

$$\bar{x} = \frac{88.22 + 100.58 + \cdots + 89.40}{100} = 95.61$$

[7] While we focus on the mean in this chapter, questions regarding variation are often just as important in practice. For instance, we would plan an event very differently if the standard deviation of runner age was 2 versus if it was 20.

The sample mean $\bar{x} = 95.61$ minutes a **point estimate** of the population mean. Suppose we take a new sample of 100 people and recompute the mean; we will probably not get the exact same answer that we got using the `run10Samp` data set. Estimates generally vary from one sample to another, and this **sampling variation** suggests our estimate may be close, but it will not be exactly equal to the parameter.

We can also estimate the average age of participants by examining the sample mean of `age`:

$$\bar{y} = \frac{59 + 32 + \cdots + 26}{100} = 35.05$$

What about generating point estimates of other **population parameters**, such as the population median or population standard deviation? Once again we might estimate parameters based on sample statistics, as shown in Table 6.5. For example, we estimate the population standard deviation for the running time using the sample standard deviation, 15.78 minutes.

`time`	estimate	parameter
mean	95.61	94.52
median	95.46	94.03
st. dev.	15.78	15.93

Table 5.8: Point estimates and parameter values for the `time` variable.

⊙ **Exercise 5.4** Suppose we want to estimate the difference in run times for men and women. If $\bar{x}_{men} = 87.65$ and $\bar{x}_{women} = 102.13$, then what would be a good point estimate for the population difference?[8]

⊙ **Exercise 5.5** If you had to provide a point estimate of the population IQR for the run time of participants, how might you make such an estimate using a sample?[9]

As mentioned in Section 6.1 point estimates are usually not exactly equal to the truth, but they get better as more data become available. We can see this by plotting a running mean from our `run10Samp` sample. A **running mean** is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence. For example, the second mean in the sequence is the average of the first two observations and the third in the sequence is the average of the first three. The running mean for the 10 mile run time in the `run10Samp` data set is shown in Figure 6.6, and it approaches the true population average, 94.52 minutes, as more data become available.

[8] We could take the difference of the two sample means: $102.13 - 87.65 = 14.48$. Men ran about 14.48 minutes faster on average in the 2012 Cherry Blossom Run.

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

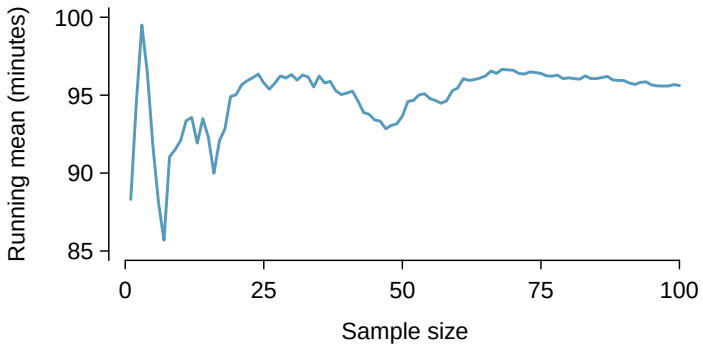


Figure 5.9: The mean computed after adding each individual to the sample. The mean tends to approach the true population average as more data become available.

Sample point estimates only approximate the population parameter, and they vary from one sample to another. If we took another simple random sample of the Cherry Blossom runners, we would find that the sample mean for the run time would be a little different. It will be useful to quantify how variable an estimate is from one sample to another. If this variability is small (i.e. the sample mean doesn't change much from one sample to another) then that estimate is probably very accurate. If it varies widely from one sample to another, then we should not expect our estimate to be very good.

5.4.2 A sampling distribution of the sample mean

From the random sample represented in `run10Samp`, we guessed the average time it takes to run 10 miles is 95.61 minutes. Suppose we take another random sample of 100 individuals and take its mean: 95.30 minutes. Suppose we took another (93.43 minutes) and another (94.16 minutes), and so on. If we do this many many times – which we can do only because we have the entire population data set – we can build up a **sampling distribution** for the sample mean when the sample size is 100, shown in Figure 6.7.

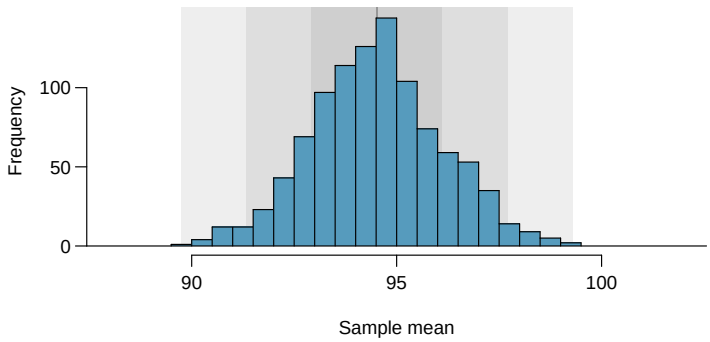


Figure 5.10: A histogram of 1000 sample means for run time, where the samples are of size $n = 100$.

In Figure 6.7 we have introduced a sampling distribution for $\bar{x}$, the average run time

for samples of size 100. The shape of the distribution looks fairly normal with 1000 samples each of size 100. Now we'll take 100,000 samples, calculate the mean of each, and plot them in a histogram to get an especially accurate depiction of the sampling distribution. This histogram is shown in the left panel of Figure 6.9.

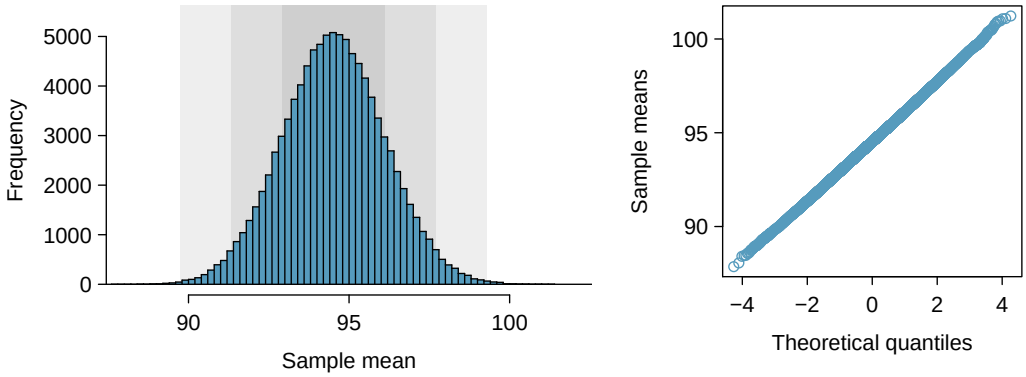


Figure 5.11: The left panel shows a histogram of the sample means for 100,000 different random samples. The right panel shows a normal probability plot of those sample means.

Does this distribution look familiar? Hopefully so! The distribution of sample means closely resembles the normal distribution (see Section 4.2.2). A normal probability plot of these sample means is shown in the right panel of Figure 6.9. Because all of the points closely fall around a straight line, we can conclude the distribution of sample means is nearly normal. This result can be explained by the Central Limit Theorem.

5.4.3 Standard error of the mean

The sampling distribution shown in Figure 6.7 is unimodal and approximately symmetric. It is also centered exactly at the true population mean: $\mu = 94.52$. Intuitively, this makes sense. The sample means should tend to "fall around" the population mean.

⊙ **Exercise 5.6** (a) Would you rather use a small sample or a large sample when estimating a parameter? Why? (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard error than a point estimate based on a larger sample?[10]

In the sample of 100 runners, the standard error of the sample mean is equal to one-tenth of the population standard deviation: $1.59 = 15.93/10$. In other words, the standard error of the sample mean based on 100 observations is equal to

$$SE_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{15.93}{\sqrt{100}} = 1.59$$

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

where σ_x is the standard deviation of the individual observations. This is no coincidence; it is a direct result of using the central limit theorem and the probability tools of Section 3.4.

⊙ **Exercise 5.7** In the sample of 100 runners, the standard deviation of the runners' ages is $s_y = 8.97$. Because the sample is simple random and consists of less than 10% of the population, the observations are independent. (a) What is the standard error of the sample mean, $\bar{y} = 35.05$ years? (b) Would you be surprised if someone told you the average age of all the runners was actually 36 years?[11]

⊙ **Exercise 5.8** (a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation of the individual observations is 10, what is our estimate of the standard error when the sample size is 100? What about when it is 400? (c) Explain how your answer to (b) mathematically justifies your intuition in part (a).[12]

[11] (a) Use the solution of Exercise (5.6) with the sample standard deviation to compute the standard error: $SE_{\bar{y}} = 8.97/\sqrt{100} = 0.90$ years. (b) It would not be surprising. Our sample is about 1 standard error from 36 years. In other words, 36 years old does not seem to be implausible given that our sample was relatively close to it. (We use the standard error to identify what is close.)

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

Chapter 6

Foundations for inference

Statistical inference is concerned primarily with understanding the quality of parameter estimates. For example, a classic inferential question is, "How sure are we that the estimated mean, $\bar{x}$, is near the true population mean, μ?" While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. We introduce these common themes in Sections 6.1-5.2 by discussing inference about the population mean, μ, and set the stage for other parameters and scenarios in Section 6.3. Some advanced considerations are discussed in Section 6.4. Understanding this chapter will make the rest of this book, and indeed the rest of statistics, seem much more familiar.

Throughout the next few sections we consider a data set called `run10`, which represents all 16,924 runners who finished the 2012 Cherry Blossom 10 mile run in Washington, DC.[1] Part of this data set is shown in Table 6.1, and the variables are described in Table 6.2.

ID	time	age	gender	state
1	92.25	38.00	M	MD
2	106.35	33.00	M	DC
3	89.33	55.00	F	VA
4	113.50	24.00	F	VA
⋮	⋮	⋮	⋮	⋮
16923	122.87	37.00	F	VA
16924	93.30	27.00	F	DC

Table 6.1: Six observations from the `run10` data set.

variable	**description**
`time`	Ten mile run time, in minutes
`age`	Age, in years
`gender`	Gender (`M` for male, `F` for female)
`state`	Home state (or country if not from the US)

Table 6.2: Variables and their descriptions for the `run10` data set.

[1] http://www.cherryblossom.org

ID	time	age	gender	state
1983	88.31	59	M	MD
8192	100.67	32	M	VA
11020	109.52	33	F	VA
⋮	⋮	⋮	⋮	⋮
1287	89.49	26	M	DC

Table 6.3: Four observations for the `run10Samp` data set, which represents a simple random sample of 100 runners from the 2012 Cherry Blossom Run.

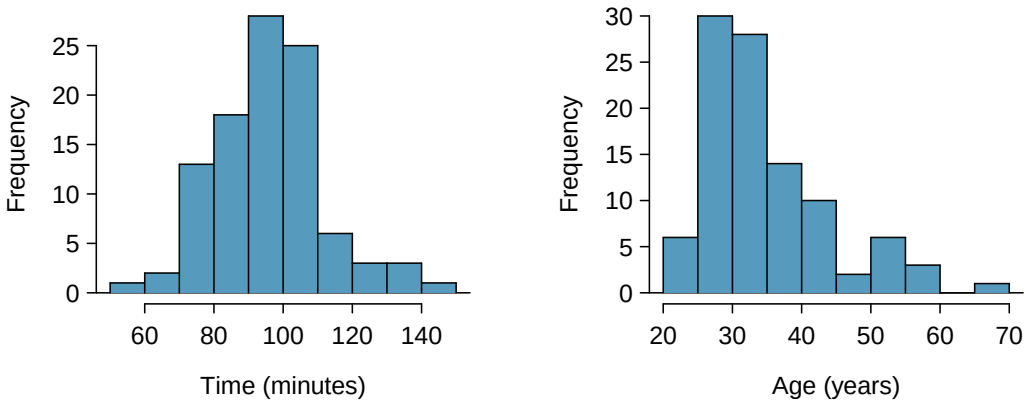


Figure 6.4: Histograms of `time` and `age` for the sample Cherry Blossom Run data. The average time is in the mid-90s, and the average age is in the mid-30s. The age distribution is moderately skewed to the right.

These data are special because they include the results for the entire population of runners who finished the 2012 Cherry Blossom Run. We took a simple random sample of this population, which is represented in Table 6.3. We will use this sample, which we refer to as the `run10Samp` data set, to draw conclusions about the entire population. This is the practice of statistical inference in the broadest sense. Two histograms summarizing the time and age variables in the `run10Samp` data set are shown in Figure 6.4.

6.1 Variability in estimates

We would like to estimate two features of the Cherry Blossom runners using the sample.

(1) How long does it take a runner, on average, to complete the 10 miles?

(2) What is the average age of the runners?

These questions may be informative for planning the Cherry Blossom Run in future years.[2] We will use $x_1, ..., x_{100}$ to represent the 10 mile time for each runner in our sample, and $y_1, ..., y_{100}$ will represent the age of each of these participants.

[2]While we focus on the mean in this chapter, questions regarding variation are often just as important in practice. For instance, we would plan an event very differently if the standard deviation of runner age was 2 versus if it was 20.

6.1.1 Point estimates

We want to estimate the **population mean** based on the sample. The most intuitive way to go about doing this is to simply take the **sample mean**. That is, to estimate the average 10 mile run time of all participants, take the average time for the sample:

$$\bar{x} = \frac{88.22 + 100.58 + \cdots + 89.40}{100} = 95.61$$

The sample mean $\bar{x} = 95.61$ minutes is called a **point estimate** of the population mean: if we can only choose one value to estimate the population mean, this is our best guess. Suppose we take a new sample of 100 people and recompute the mean; we will probably not get the exact same answer that we got using the `run10Samp` data set. Estimates generally vary from one sample to another, and this **sampling variation** suggests our estimate may be close, but it will not be exactly equal to the parameter.

We can also estimate the average age of participants by examining the sample mean of `age`:

$$\bar{y} = \frac{59 + 32 + \cdots + 26}{100} = 35.05$$

What about generating point estimates of other **population parameters**, such as the population median or population standard deviation? Once again we might estimate parameters based on sample statistics, as shown in Table 6.5. For example, we estimate the population standard deviation for the running time using the sample standard deviation, 15.78 minutes.

`time`	estimate	parameter
mean	95.61	94.52
median	95.46	94.03
st. dev.	15.78	15.93

Table 6.5: Point estimates and parameter values for the `time` variable.

⊙ **Exercise 6.1** Suppose we want to estimate the difference in run times for men and women. If $\bar{x}_{men} = 87.65$ and $\bar{x}_{women} = 102.13$, then what would be a good point estimate for the population difference?[3]

⊙ **Exercise 6.2** If you had to provide a point estimate of the population IQR for the run time of participants, how might you make such an estimate using a sample?[4]

6.1.2 Point estimates are not exact

Estimates are usually not exactly equal to the truth, but they get better as more data become available. We can see this by plotting a running mean from our `run10Samp` sample. A **running mean** is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence. For example, the second mean in the sequence is the average of the first two observations and the third in the

[3] We could take the difference of the two sample means: $102.13 - 87.65 = 14.48$. Men ran about 14.48 minutes faster on average in the 2012 Cherry Blossom Run.

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

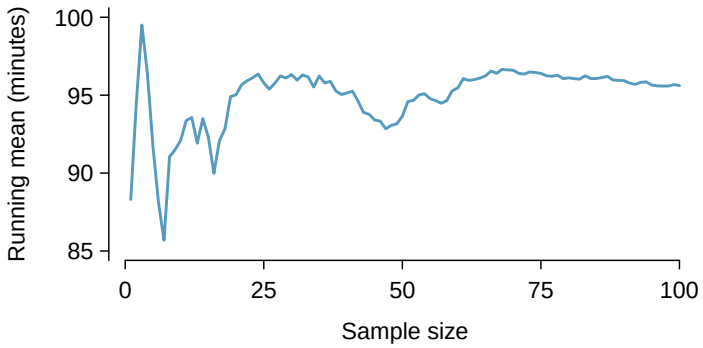


Figure 6.6: The mean computed after adding each individual to the sample. The mean tends to approach the true population average as more data become available.

sequence is the average of the first three. The running mean for the 10 mile run time in the `run10Samp` data set is shown in Figure 6.6, and it approaches the true population average, 94.52 minutes, as more data become available.

Sample point estimates only approximate the population parameter, and they vary from one sample to another. If we took another simple random sample of the Cherry Blossom runners, we would find that the sample mean for the run time would be a little different. It will be useful to quantify how variable an estimate is from one sample to another. If this variability is small (i.e. the sample mean doesn't change much from one sample to another) then that estimate is probably very accurate. If it varies widely from one sample to another, then we should not expect our estimate to be very good.

6.1.3 Standard error of the mean

From the random sample represented in `run10Samp`, we guessed the average time it takes to run 10 miles is 95.61 minutes. Suppose we take another random sample of 100 individuals and take its mean: 95.30 minutes. Suppose we took another (93.43 minutes) and another (94.16 minutes), and so on. If we do this many many times – which we can do only because we have the entire population data set – we can build up a **sampling distribution** for the sample mean when the sample size is 100, shown in Figure 6.7.

Sampling distribution
The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from such a distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

The sampling distribution shown in Figure 6.7 is unimodal and approximately symmetric. It is also centered exactly at the true population mean: $\mu = 94.52$. Intuitively, this makes sense. The sample means should tend to "fall around" the population mean.

We can see that the sample mean has some variability around the population mean, which can be quantified using the standard deviation of this distribution of sample means: $\sigma_{\bar{x}} = 1.59$. The standard deviation of the sample mean tells us how far the typical estimate

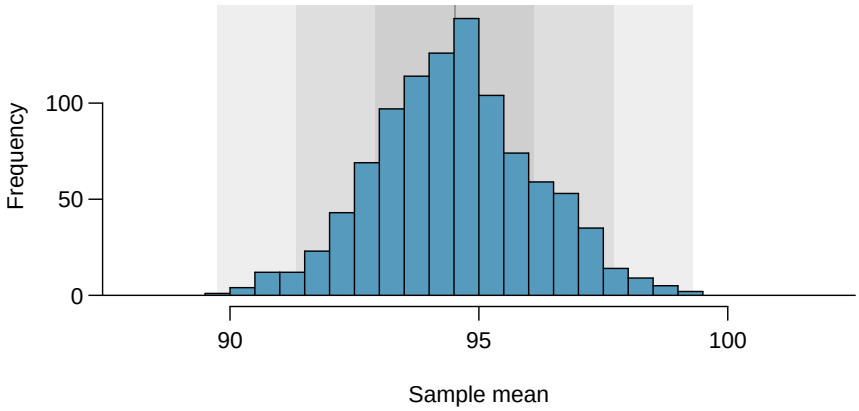


Figure 6.7: A histogram of 1000 sample means for run time, where the samples are of size $n = 100$.

is away from the actual population mean, 94.52 minutes. It also describes the typical **error** of the point estimate, and for this reason we usually call this standard deviation the **standard error (SE)** of the estimate.

SE
standard error

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

When considering the case of the point estimate $\bar{x}$, there is one problem: there is no obvious way to estimate its standard error from a single sample. However, statistical theory provides a helpful tool to address this issue.

⊙ **Exercise 6.3** (a) Would you rather use a small sample or a large sample when estimating a parameter? Why? (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard error than a point estimate based on a larger sample?[5]

In the sample of 100 runners, the standard error of the sample mean is equal to one-tenth of the population standard deviation: $1.59 = 15.93/10$. In other words, the standard error of the sample mean based on 100 observations is equal to

$$SE_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{15.93}{\sqrt{100}} = 1.59$$

where σ_x is the standard deviation of the individual observations. This is no coincidence. We can show mathematically that this equation is correct when the observations are independent using the probability tools of Section 3.4.

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

Computing SE for the sample mean

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

$$SE = \frac{\sigma}{\sqrt{n}} \tag{6.4}$$

A reliable method to ensure sample observations are independent is to conduct a simple random sample consisting of less than 10% of the population.

There is one subtle issue of Equation ((6.4)): the population standard deviation is typically unknown. You might have already guessed how to resolve this problem: we can use the point estimate of the standard deviation from the sample. This estimate tends to be sufficiently good when the sample size is at least 30 and the population distribution is not strongly skewed. Thus, we often just use the sample standard deviation s instead of σ. When the sample size is smaller than 30, we will need to use a method to account for extra uncertainty in the standard error. If the skew condition is not met, a larger sample is needed to compensate for the extra skew. These topics are further discussed in Section 5.2.

⊙ **Exercise 6.5** In the sample of 100 runners, the standard deviation of the runners' ages is $s_y = 8.97$. Because the sample is simple random and consists of less than 10% of the population, the observations are independent. (a) What is the standard error of the sample mean, $\bar{y} = 35.05$ years? (b) Would you be surprised if someone told you the average age of all the runners was actually 36 years?[6]

⊙ **Exercise 6.6** (a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation of the individual observations is 10, what is our estimate of the standard error when the sample size is 100? What about when it is 400? (c) Explain how your answer to (b) mathematically justifies your intuition in part (a).[7]

6.1.4 Basic properties of point estimates

We achieved three goals in this section. First, we determined that point estimates from a sample may be used to estimate population parameters. We also determined that these point estimates are not exact: they vary from one sample to another. Lastly, we quantified the uncertainty of the sample mean using what we call the standard error, mathematically represented in Equation ((6.4)). While we could also quantify the standard error for other estimates – such as the median, standard deviation, or any other number of statistics – we will postpone these extensions until later chapters or courses.

[6] (a) Use Equation ((6.4)) with the sample standard deviation to compute the standard error: $SE_{\bar{y}} = 8.97/\sqrt{100} = 0.90$ years. (b) It would not be surprising. Our sample is about 1 standard error from 36 years. In other words, 36 years old does not seem to be implausible given that our sample was relatively close to it. (We use the standard error to identify what is close.)

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

6.2 Confidence intervals

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

In this section and in Section 8.1, we will emphasize the special case where the point estimate is a sample mean and the parameter is the population mean. In Section 6.3, we generalize these methods for a variety of point estimates and population parameters that we will encounter in Chapter **??** and beyond.

6.2.1 Capturing the population parameter

A plausible range of values for the population parameter is called a **confidence interval**.

Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

⊙ **Exercise 6.7** If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[8]

6.2.2 An approximate 95% confidence interval

Our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate. The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

The standard error represents the standard deviation associated with the estimate, and roughly 95% of the time the estimate will be within 2 standard errors of the parameter. If the interval spreads out 2 standard errors from the point estimate, we can be roughly 95% **confident** that we have captured the true parameter:

$$\text{point estimate} \pm 2 \times SE \tag{6.8}$$

But what does "95% confident" mean? Suppose we took many samples and built a confidence interval from each sample using Equation (7.1.2). Then about 95% of those intervals would contain the actual mean, μ. Figure 7.1 shows this process with 25 samples, where 24 of the resulting confidence intervals contain the average time for all the runners, $\mu = 94.52$ minutes, and one does not.

⊙ **Exercise 6.9** In Figure 7.1, one interval does not contain 94.52 minutes. Does this imply that the mean cannot be 94.52? [9]

[8] 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.

[9] Just as some observations occur more than 2 standard deviations from the mean, some point estimates will be more than 2 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are impossible.

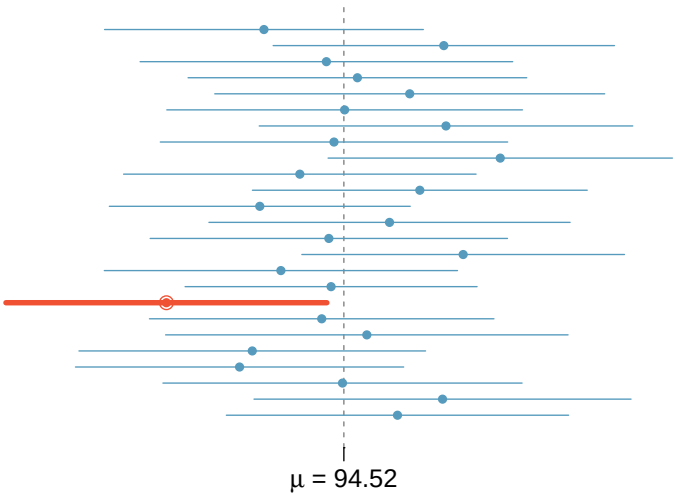


Figure 6.8: Twenty-five samples of size $n = 100$ were taken from the `run10` data set. For each sample, a confidence interval was created to try to capture the average 10 mile time for the population. Only 1 of these 25 intervals did not capture the true mean, $\mu = 94.52$ minutes.

The rule where about 95% of observations are within 2 standard deviations of the mean is only approximately true. However, it holds very well for the normal distribution. As we will soon see, the mean tends to be normally distributed when the sample size is sufficiently large.

Example 6.10 If the sample mean of times from `run10Samp` is 95.61 minutes and the standard error, as estimated using the sample standard deviation, is 1.58 minutes, what would be an approximate 95% confidence interval for the average 10 mile time of all runners in the race? Apply the standard error calculated using the sample standard deviation ($SE = \frac{15.78}{\sqrt{100}} = 1.58$), which is how we usually proceed since the population standard deviation is generally unknown.

We apply Equation (7.1.2):

$$95.61 \ \pm \ 2 \times 1.58 \quad \rightarrow \quad (92.45, 98.77)$$

Based on these data, we are about 95% confident that the average 10 mile time for all runners in the race was larger than 92.45 but less than 98.77 minutes. Our interval extends out 2 standard errors from the point estimate, $\bar{x}$.

Exercise 6.11 The sample data suggest the average runner's age is about 35.05 years with a standard error of 0.90 years (estimated using the sample standard deviation, 8.97). What is an approximate 95% confidence interval for the average age of all of the runners?[10]

[10] Again apply Equation (7.1.2): $35.05 \pm 2 \times 0.90 \rightarrow (33.25, 36.85)$. We interpret this interval as follows: We are about 95% confident the average age of all participants in the 2012 Cherry Blossom Run was between 33.25 and 36.85 years.

6.2.3 A sampling distribution of the sample mean

In Section 6.1.3, we introduced a sampling distribution for $\bar{x}$, the average run time for samples of size 100. We examined this distribution earlier in Figure 6.7. Now we'll take 100,000 samples, calculate the mean of each, and plot them in a histogram to get an especially accurate depiction of the sampling distribution. This histogram is shown in the left panel of Figure 6.9.

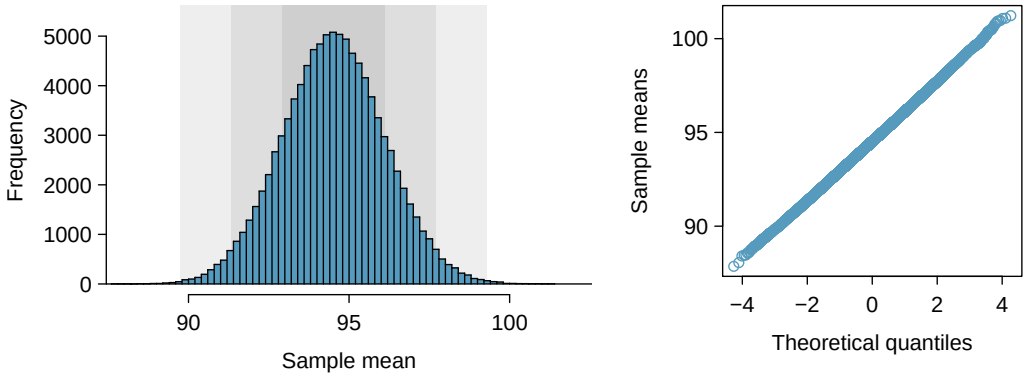


Figure 6.9: The left panel shows a histogram of the sample means for 100,000 different random samples. The right panel shows a normal probability plot of those sample means.

Does this distribution look familiar? Hopefully so! The distribution of sample means closely resembles the normal distribution (see Section 4.2.2). A normal probability plot of these sample means is shown in the right panel of Figure 6.9. Because all of the points closely fall around a straight line, we can conclude the distribution of sample means is nearly normal. This result can be explained by the Central Limit Theorem.

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

We will apply this informal version of the Central Limit Theorem for now, and discuss its details further in Section 5.2.

The choice of using 2 standard errors in Equation (7.1.2) was based on our general guideline that roughly 95% of the time, observations are within two standard deviations of the mean. Under the normal model, we can make this more accurate by using 1.96 in place of 2.

$$\text{point estimate} \pm 1.96 \times SE \tag{6.12}$$

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

6.2.4 Changing the confidence level

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

The 95% confidence interval structure provides guidance in how to make intervals with new confidence levels. Below is a general 95% confidence interval for a point estimate that comes from a nearly normal distribution:

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{6.13}$$

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 deviations of the parameter about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

⊙ **Exercise 6.14** If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[11]

To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58. Exercise (6.14) highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean. This approach – using the Z scores in the normal model to compute confidence levels – is appropriate when $\bar{x}$ is associated with a normal distribution with mean μ and standard deviation $SE_{\bar{x}}$. Thus, the formula for a 99% confidence interval is

$$\bar{x} \ \pm \ 2.58 \times SE_{\bar{x}} \tag{6.15}$$

The normal approximation is crucial to the precision of these confidence intervals. Section 5.2 provides a more detailed discussion about when the normal model can safely be applied. When the normal model is not a good fit, we will use alternative distributions that better characterize the sampling distribution.

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

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

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The distribution of sample observations is not strongly skewed.

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

[11] 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 7.5.) To determine this probability, look up -2.58 and 2.58 in the normal probability table (0.0049 and 0.9951). Thus, there is a $0.9951 - 0.0049 \approx 0.99$ probability that the unobserved random variable X will be within 2.58 standard deviations of μ.

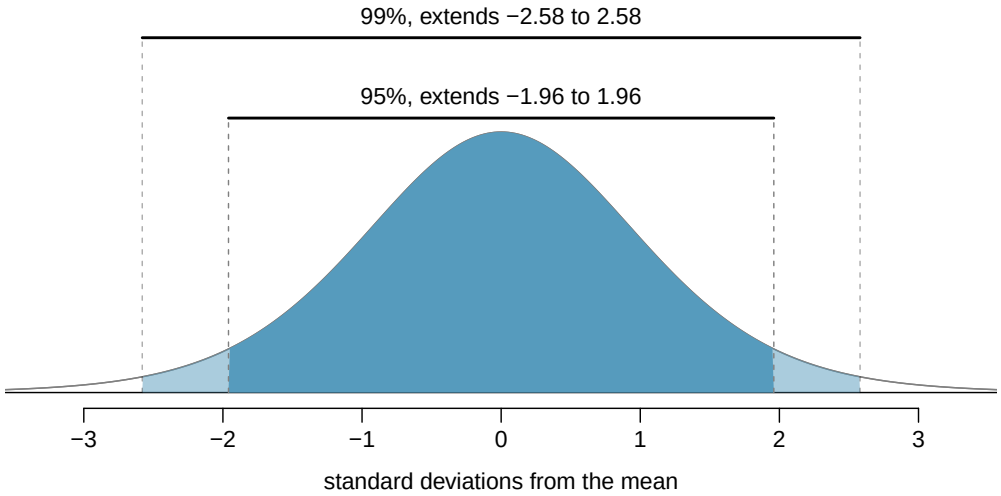


Figure 6.10: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

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

TIP: How to verify sample observations are independent
Observations in a simple random sample consisting of less than 10% of the population are independent.

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

⊙ **Exercise 6.16** Create a 99% confidence interval for the average age of all runners in the 2012 Cherry Blossom Run. The point estimate is $\bar{y} = 35.05$ and the standard error is $SE_{\bar{y}} = 0.90$.[12]

[12] The observations are independent (simple random sample, < 10% of the population), the sample size is at least 30 ($n = 100$), and the distribution is only slightly skewed (Figure 6.4); the normal approximation and estimate of SE should be reasonable. Apply the 99% confidence interval formula: $\bar{y} \pm 2.58 \times SE_{\bar{y}} \rightarrow (32.7, 37.4)$. We are 99% confident that the average age of all runners is between 32.7 and 37.4 years.

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

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

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

Figure 7.5 provides a picture of how to identify $z^\star$ based on a confidence level. We select $z^\star$ so that the area between -$z^\star$ and $z^\star$ in the normal model corresponds to the confidence level.

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

⊙ **Exercise 6.17** Use the data in Exercise (7.14) to create a 90% confidence interval for the average age of all runners in the 2012 Cherry Blossom Run.[13]

6.2.5 Interpreting confidence intervals

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

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

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

Another especially important consideration of confidence intervals is that they *only try to capture the population parameter*. Our intervals say nothing about the confidence of capturing individual observations, a proportion of the observations, or about capturing point estimates. Confidence intervals only attempt to capture population parameters.

6.2.6 Nearly normal population with known SD (special topic)

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

(1) The observations are independent.

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

(3) The data distribution is not strongly skewed.

[13] We first find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal model, $N(\mu = 0, \sigma = 1)$. We can look up -$z^\star$ in the normal probability table by looking for a lower tail of 5% (the other 5% is in the upper tail), thus $z^\star = 1.65$. The 90% confidence interval can then be computed as $\bar{y} \pm 1.65 \times SE_{\bar{y}} \rightarrow (33.6, 36.5)$. (We had already verified conditions for normality and the standard error.) That is, we are 90% confident the average age is larger than 33.6 but less than 36.5 years.

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

Example 6.18 The heights of male seniors in high school closely follow a normal distribution $N(\mu = 70.43, \sigma = 2.73)$, where the units are inches.[14] If we randomly sampled the heights of five male seniors, what distribution should the sample mean follow?

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

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

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

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

Example 6.19 Is there a connection between height and popularity in high school? Many students may suspect as much, but what do the data say? Suppose the top 5 nominees for prom king at a high school have an average height of 71.8 inches. Does this provide strong evidence that these seniors' heights are not representative of all male seniors at their high school?

If these five seniors are height-representative, then their heights should be like a random sample from the distribution given in Example (6.18), $N(\mu = 70.43, \sigma = 2.73)$, and the sample mean should follow $N(\mu = 70.43, \sigma/\sqrt{n} = 1.22)$. Formally we are conducting what is called a *hypothesis test*, which we will discuss in greater detail during the next section. We are weighing two possibilities:

H_0: The prom king nominee heights are representative; $\bar{x}$ will follow a normal distribution with mean 70.43 inches and standard error 1.22 inches.

H_A: The heights are not representative; we suspect the mean height is different from 70.43 inches.

If there is strong evidence that the sample mean is not from the normal distribution provided in H_0, then that suggests the heights of prom king nominees are not a simple random sample (i.e. H_A is true). We can look at the Z score of the sample mean to

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

tell us how unusual our sample is. If H_0 is true:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70.43}{1.22} = 1.12$$

A Z score of just 1.12 is not very unusual (we typically use a threshold of ± 2 to decide what is unusual), so there is not strong evidence against the claim that the heights are representative. This does not mean the heights are actually representative, only that this very small sample does not necessarily show otherwise.

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

6.3 Inference for other estimators

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

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

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

6.3.1 Confidence intervals for nearly normal point estimates

In Section 7, we used the point estimate $\bar{x}$ with a standard error $SE_{\bar{x}}$ to create a 95% confidence interval for the population mean:

$$\bar{x} \pm 1.96 \times SE_{\bar{x}} \tag{6.20}$$

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

General confidence interval for the normal sampling distribution case
A confidence interval based on an unbiased and nearly normal point estimate is

$$\text{point estimate} \pm z^\star SE \tag{6.21}$$

where $z^\star$ is selected to correspond to the confidence level, and SE represents the standard error. The value $z^\star SE$ is called the *margin of error*.

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

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

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

Example 6.22 In Exercise (6.1) on page 161, we computed a point estimate for the average difference in run times between men and women: $\bar{x}_{women} - \bar{x}_{men} = 14.48$ minutes. This point estimate is associated with a nearly normal distribution with standard error $SE = 2.78$ minutes. What is a reasonable 95% confidence interval for the difference in average run times?

The normal approximation is said to be valid, so we apply Equation ((6.21)):

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 14.48 \pm 1.96 \times 2.78 \quad \rightarrow \quad (9.03, 19.93)$$

Thus, we are 95% confident that the men were, on average, between 9.03 and 19.93 minutes faster than women in the 2012 Cherry Blossom Run. That is, the actual average difference is plausibly between 9.03 and 19.93 minutes with 95% confidence.

Example 6.23 Does Example (6.22) guarantee that if a husband and wife both ran in the race, the husband would run between 9.03 and 19.93 minutes faster than the wife?

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

Exercise 6.24 What $z^\star$ would be appropriate for a 99% confidence level? For help, see Figure 7.5 on page 186.[15]

Exercise 6.25 The proportion of men in the `run10Samp` sample is $\hat{p} = 0.45$. This sample meets certain conditions that ensure $\hat{p}$ will be nearly normal, and the standard error of the estimate is $SE_{\hat{p}} = 0.05$. Create a 90% confidence interval for the proportion of participants in the 2012 Cherry Blossom Run who are men.[16]

[15] We seek $z^\star$ such that 99% of the area under the normal curve will be between the Z scores -$z^\star$ and $z^\star$. Because the remaining 1% is found in the tails, each tail has area 0.5%, and we can identify -$z^\star$ by looking up 0.0050 in the normal probability table: $z^\star = 2.58$. See also Figure 7.5 on page 186.

[16] We use $z^\star = 1.65$ (see Exercise (7.15) on page 188), and apply the general confidence interval formula:

$$\hat{p} \pm z^\star SE_{\hat{p}} \quad \rightarrow \quad 0.45 \pm 1.65 \times 0.05 \quad \rightarrow \quad (0.3675, 0.5325)$$

Thus, we are 90% confident that between 37% and 53% of the participants were men.

6.3.2 Hypothesis testing for nearly normal point estimates

Just as the confidence interval method works with many other point estimates, we can generalize our hypothesis testing methods to new point estimates. Here we only consider the p-value approach, introduced in Section **??**, since it is the most commonly used technique and also extends to non-normal cases.

Hypothesis testing using the normal model

1. First write the hypotheses in plain language, then set them up in mathematical notation.
2. Identify an appropriate point estimate of the parameter of interest.
3. Verify conditions to ensure the standard error estimate is reasonable and the point estimate is nearly normal and unbiased.
4. Compute the standard error. Draw a picture depicting the distribution of the estimate under the idea that H_0 is true. Shade areas representing the p-value.
5. Using the picture and normal model, compute the *test statistic* (Z score) and identify the p-value to evaluate the hypotheses. Write a conclusion in plain language.

⊙ **Exercise 6.26** A drug called sulphinpyrazone was under consideration for use in reducing the death rate in heart attack patients. To determine whether the drug was effective, a set of 1,475 patients were recruited into an experiment and randomly split into two groups: a control group that received a placebo and a treatment group that received the new drug. What would be an appropriate null hypothesis? And the alternative?[17]

We can formalize the hypotheses from Exercise (6.26) by letting $p_{control}$ and $p_{treatment}$ represent the proportion of patients who died in the control and treatment groups, respectively. Then the hypotheses can be written as

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

or equivalently,

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

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

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

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

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

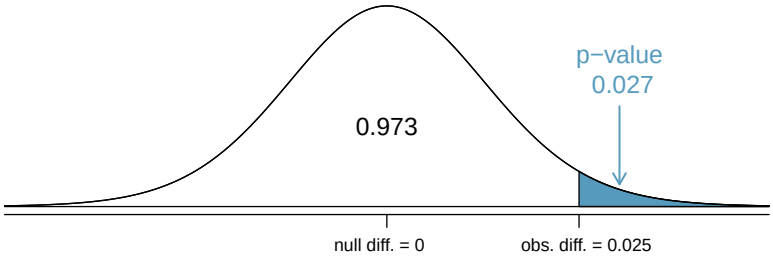


Figure 6.11: The distribution of the sample difference if the null hypothesis is true.

Example 6.27 We want to evaluate the hypothesis setup from Exericse (6.26) using data from the actual study.[18] In the control group, 60 of 742 patients died. In the treatment group, 41 of 733 patients died. The sample difference in death rates can be summarized as

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

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

We would like to identify the p-value to evaluate the hypotheses. If the null hypothesis is true, then the point estimate would have come from a nearly normal distribution, like the one shown in Figure 6.11. The distribution is centered at zero since $p_{control} - p_{treatment} = 0$ under the null hypothesis. Because a large positive difference provides evidence against the null hypothesis and in favor of the alternative, the upper tail has been shaded to represent the p-value. We need not shade the lower tail since this is a one-sided test: an observation in the lower tail does not support the alternative hypothesis.

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

$$Z = \frac{\text{point estimate} - \text{null value}}{SE_{\text{point estimate}}} = \frac{0.025 - 0}{0.013} = 1.92 \tag{6.28}$$

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

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

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

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

The Z score in Equation ((6.28)) is called a **test statistic**. In most hypothesis tests, a test statistic is a particular data summary that is especially useful for computing the p-value and evaluating the hypothesis test. In the case of point estimates that are nearly normal, the test statistic is the Z score.

Test statistic
A *test statistic* is a special summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. When a point estimate is nearly normal, we use the Z score of the point estimate as the test statistic. In later chapters we encounter situations where other test statistics are helpful.

6.3.3 Non-normal point estimates

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

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

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

6.3.4 When to retreat

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

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

Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a statistician.[19] The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings.

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

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

6.4 Sample size and power (special topic)

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

6.4.1 Finding a sample size for a certain margin of error

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

Example 6.29 Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg. How large of a sample is necessary to estimate the average systolic blood pressure with a margin of error of 4 mmHg using a 95% confidence level?

First, we frame the problem carefully. Recall that the margin of error is the part we add and subtract from the point estimate when computing a confidence interval. The margin of error for a 95% confidence interval estimating a mean can be written as

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

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

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

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

for n, we substitute 25 for $\sigma_{employee}$:

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

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

A potentially controversial part of Example (6.29) is the use of the U.S. standard deviation for the employee standard deviation. Usually the standard deviation is not known. In such cases, it is reasonable to review scientific literature or market research to make an educated guess about the standard deviation.

Identify a sample size for a particular margin of error

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

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

where $z^\star$ is chosen to correspond to the desired confidence level, and σ is the standard deviation associated with the population. Solve for the sample size, n.

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

6.4.2 Power and the Type 2 Error rate

Consider the following two hypotheses:

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

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

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

Example 6.30 Suppose the actual employee average is 132 and we take a sample of 100 individuals. Then the true sampling distribution of $\bar{x}$ is approximately $N(132, 2.5)$ (since $SE = \frac{25}{\sqrt{100}} = 2.5$). What is the probability of successfully rejecting the null hypothesis?

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

Step 1. The null distribution could be represented by $N(130, 2.5)$, the same standard deviation as the true distribution but with the null value as its center. Then we can find the two tail areas by identifying the Z score corresponding to the 2.5% tails (± 1.96), and solving for x in the Z score equation:

$$-1.96 = Z_1 = \frac{x_1 - 130}{2.5} \qquad +1.96 = Z_2 = \frac{x_2 - 130}{2.5}$$
$$x_1 = 125.1 \qquad x_2 = 134.9$$

(An equally valid approach is to recognize that x_1 is $1.96 \times SE$ below the mean and x_2 is $1.96 \times SE$ above the mean to compute the values.) Figure 6.12 shows the null distribution on the left with these two dotted cutoffs.

Step 2. Next, we compute the probability of rejecting H_0 if $\bar{x}$ actually came from $N(132, 2.5)$. This is the same as finding the two shaded tails for the second distribution in Figure 6.12. We use the Z score method:

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

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

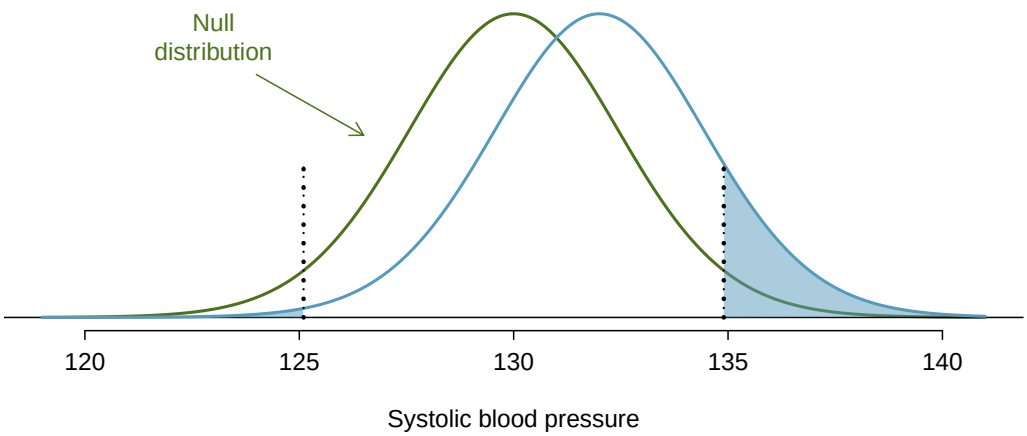


Figure 6.12: The sampling distribution of $\bar{x}$ under two scenarios. Left: $N(130, 2.5)$. Right: $N(132, 2.5)$, and the shaded areas in this distribution represent the power of the test.

The probability of rejecting the null hypothesis is called the **power**. The power varies depending on what we suppose the truth might be. In Example (6.30), the difference

between the null value and the supposed true mean was relatively small, so the power was also small: only 0.126. However, when the truth is far from the null value, where we use the standard error as a measure of what is far, the power tends to increase.

⊙ **Exercise 6.31** Suppose the true sampling distribution of $\bar{x}$ is centered at 140. That is, $\bar{x}$ comes from $N(140, 2.5)$. What would the power be under this scenario? It may be helpful to draw $N(140, 2.5)$ and shade the area representing power on Figure 6.12; use the same cutoff values identified in Example (6.30).[20]

⊙ **Exercise 6.32** If the power of a test is 0.979 for a particular mean, what is the Type 2 Error rate for this mean?[21]

⊙ **Exercise 6.33** Provide an intuitive explanation for why we are more likely to reject H_0 when the true mean is further from the null value.[22]

6.4.3 Statistical significance versus practical significance

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

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

The role of a statistician in conducting a study often includes planning the size of the study. The statistician might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain some reasonable estimate for the standard deviation. With these important pieces of information, she would choose a sufficiently large sample size so that the power for the meaningful difference is perhaps 80% or 90%. While larger sample sizes may still be used, she might advise against using them in some cases, especially in sensitive areas of research.

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

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

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

Chapter 7

Confidence intervals

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

Confidence Intervals
A plausible range of values for a parameter is called a **confidence interval**.

7.1 Introduction

7.1.1 Capturing the population parameter

Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish. If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter. The idea is to create an interval centered around the point estimate that will capture the parameter with a certain amount of confidence.

⊙ **Exercise 7.1** If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[1]

7.1.2 Constructing an approximate $(1-\alpha)\%$ confidence interval

Since our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate. The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

[1] 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.

Recall that the standard error represents the standard deviation associated with the estimate. Roughly $(1-\alpha)\%$ of the time the estimate will be within a certain number of fixed standard errors of the parameter. $(1-\alpha)$ is a fixed number that can be found out using a reference distribution such as the standard normal distribution. If an interval spreads out this fixed number of standard errors from the point estimate, we can be approximately $(1-\alpha)\%$ **confident** that we have captured the true parameter [2].

Basic skeleton of a confidence interval

$$\text{estimator} \pm \underbrace{\begin{pmatrix}\text{value from a}\\ \text{reference distribution}\end{pmatrix} \times \begin{pmatrix}\text{standard error of}\\ \text{the estimate}\end{pmatrix}}_{\text{margin of error}} \tag{7.2}$$

Although the construction of a confidence interval has essentially the same basic structure, the values that we substitute into 7.1.2 depend on the situation. The value from the reference distribution multiplied by the standard error of the estimate is called the **margin of error**. Recall that a confidence interval is a range of values. We get the upper bound of the range by adding the margin of error to the estimator and we get the lower bound by subtracting the margin of error from the estimator.

7.1.3 Interpreting an approximate $(1-\alpha)\%$ confidence interval

In Section 7.1.2 we learnt the basic procedure behind creating a $(1-\alpha)\%$ confidence interval. But what does a "$(1-\alpha)\%$ confident" mean? Let's consider the specific case where $\alpha = 0.05$. $1-\alpha = 1-0.05 = 0.95$ which means that interested in interpreting a 95% confidence interval. Suppose we took many samples and built a confidence interval from each sample using the appropriate values in Equation (7.1.2). Then approximately 95% of those intervals would contain the actual mean, μ. Figure 7.1 shows this process with 25 samples.

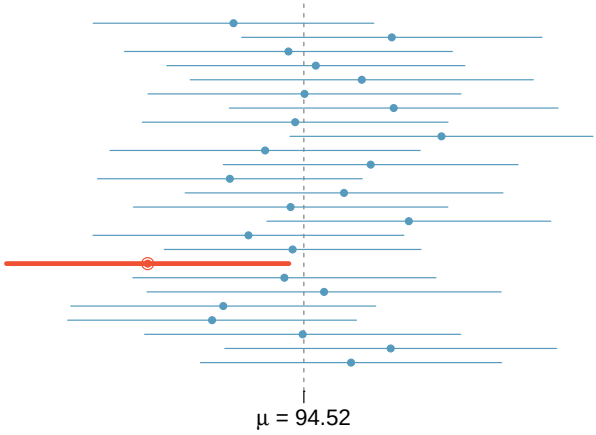


Figure 7.1: 25 samples of size $n = 100$ were taken from the `run10` data set. A 95% confidence interval for mean time was created with each sample. Only 1 interval did not capture the true mean which is $\mu = 94.52$ min.

[2] This notation might appear strange, but we are in fact using standard notation which will make more sense when we study hypothesis tests in Chapter 8.

In the example in Figure 7.1 we see that 24 out of the 25 confidence intervals (each of confidence level 95%) captured the parameter of interest which is $\mu = 94.52$ and one does not. Notice how $0.95 \times 25 \approx 24$. If we constructed 1,000 of these 95% confidence intervals then we would expect approximately 95% of them (which is $0.95 \times 1,000 = 950$) to capture the mean. In general suppose we construct a certain number of C% confidence intervals for the same population. Let's label this number of confidence intervals constructed as M. Approximately C% of these M confidence intervals will capture the population parameter. This is the formal interpretation of confidence intervals involving repeated samples.

Interpreting Confidence Intervals (Formal)
Suppose we construct a $(1 - \alpha)$% confidence interval for some parameter. In repeated sampling, we are $(1 - \alpha)$% confident that approximately $(1 - \alpha)$% of the intervals constructed using the procedure that we use will capture the target parameter.

However we can use a more intuitive interpretation of confidence intervals which is the following

Interpreting Confidence Intervals (Intuitive)
Suppose we construct a $(1 - \alpha)$% confidence interval for some parameter. Then we are $(1 - \alpha)$% confident that our target parameter is inside the interval that we constructed.

⊙ **Exercise 7.3** In Figure 7.1, one interval does not contain 94.52 minutes. Does this imply that the mean cannot be 94.52? [3]

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. We reiterate that the correct interpretation is: "We are $(1 - \alpha)$% confident that the population parameter lies between (lower bound of interval) and (upper bound of interval)". *Incorrect* language might try to describe the confidence interval as capturing the population parameter with a certain probability. For example it is wrong to state that we there is a $(1 - \alpha)$% probability that the interval captures the parameter. This is one of the most common errors: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval. The term *probability* refers to random events in which outcomes can be different when an experiment is repeated. The upper bound and lower bound of a confidence interval are fixed once they have been calculated. They are not random. As mentioned earlier in Section 6.1 the population parameter of interest is also fixed. Either the parameter is inside the calculated interval or it isn't and there is no randomness involved; hence we can not interpret a confidence interval in terms of probability.

Another especially important consideration of confidence intervals is that they *only try to capture the population parameter*. Our intervals say nothing about the confidence of capturing individual observations, a proportion of the observations, or about capturing point estimates. Confidence intervals only attempt to capture population parameters.

[3] Just as some observations occur more than 2 standard deviations from the mean, some point estimates will be more than 2 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are impossible.

7.1.4 Finding values from the relevant reference distribution

When we create a confidence interval we attach a certain amount of confidence to it. The amount of confidence we associate with it depends on a value taken from a reference distribution as mentioned in Section 7.1.2. The reference distribution we use depends on the information provided. Certain schools of thought like to make the distinction between small and large samples but we will not be doing so. We will keep things simple (as well as more faithful to the theory).

Summary of reference distributions to use for constructing confidence intervals

When σ is known	$\longrightarrow$	Use the Z distribution
When σ is not known	$\longrightarrow$	Use the t distribution
For proportions	$\longrightarrow$	Use the Z distribution

7.1.4.1 When σ is known and for proportions

When σ is known our reference distribution will be the standard normal distribution. Recall that the letter Z is associated with the standard normal distribution. The value from standard normal distribution that we require for constructing our confidence interval is labelled $z_{\alpha/2}$. The subscript $\frac{\alpha}{2}$ refers to a tail probability in our reference distribution which represents the most extreme cases. When we want to construct a $(1-\alpha)\%$ confidence interval, we can calculate $\alpha/2$ to determine the tail probabilities of the standard normal distribution and then figure out the value of $z_{\alpha/2}$ that corresponds to these extreme probabilities. The area between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ will correspond to a probability of $(1-\alpha)\%$. Figure 7.2 illustrates this procedure.

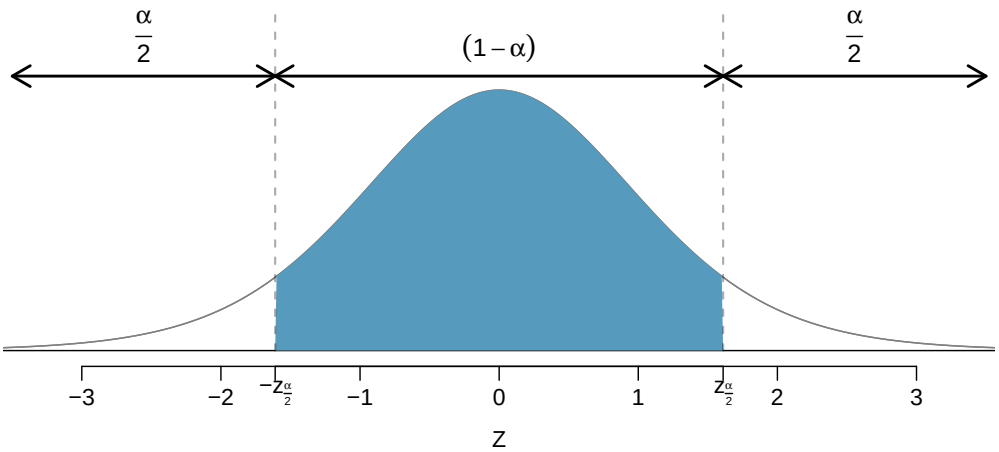


Figure 7.2: The area between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ corresponds to a $(1-\alpha)\%$ confidence interval on a proportion or on the mean when σ is known

Let's do several examples with real values.

Example 7.4 Find $z_{\alpha/2}$ for a a 95% confidence interval when σ is known or for

proportions.

This means that we have

$$1 - \alpha = 0.95 \tag{7.5}$$

$$\frac{\alpha}{2} = 0.025 \tag{7.6}$$

This means that the area in each tail is 0.025. From Appendix A we see that this corresponds to a value of $z = 1.96$. Hence $z_{\alpha/2=1.96}$ for a 95% confidence interval when $\sigma isknown$ or on proportions. This is illustrated in Figure 7.3.

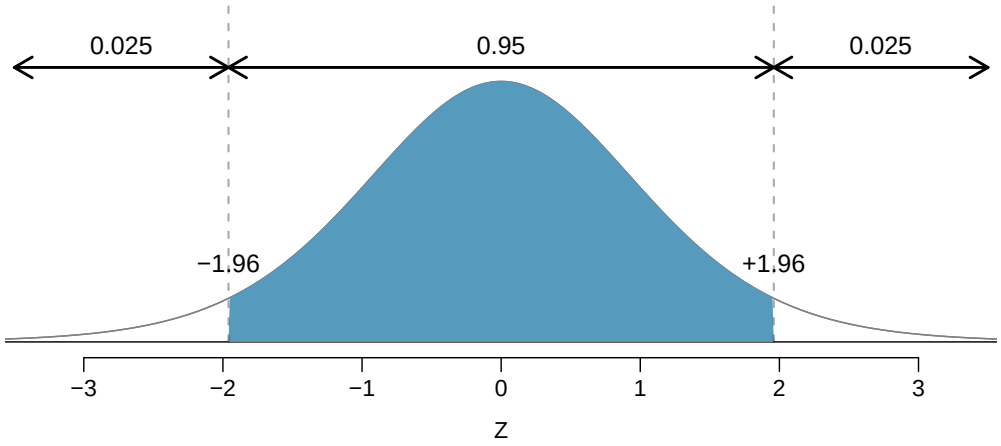


Figure 7.3: The area between $-z_{\alpha/2} = -1.96$ and $+z_{\alpha/2} = +1.96$ corresponds to a 95% confidence interval on a proportion or on the mean when σ is known

Example 7.7 Find $z_{\alpha/2}$ for a a 99% confidence interval when σ is known or for proportions.

We are essentially repeating the procedure in Example (7.4) for a 99% confidence interval when σ is known or for proportions. This time we have

$$1 - \alpha = 0.99 \tag{7.8}$$

$$\frac{\alpha}{2} = 0.005 \tag{7.9}$$

Now the area we look for in each tail is 0.025. From Appendix A we see that this corresponds to a value of $z = 2.58$. Hence $z_{\alpha/2=2.58}$ for a 99% confidence interval on proportions or when σ is known. This is illustrated in Figure 7.4.

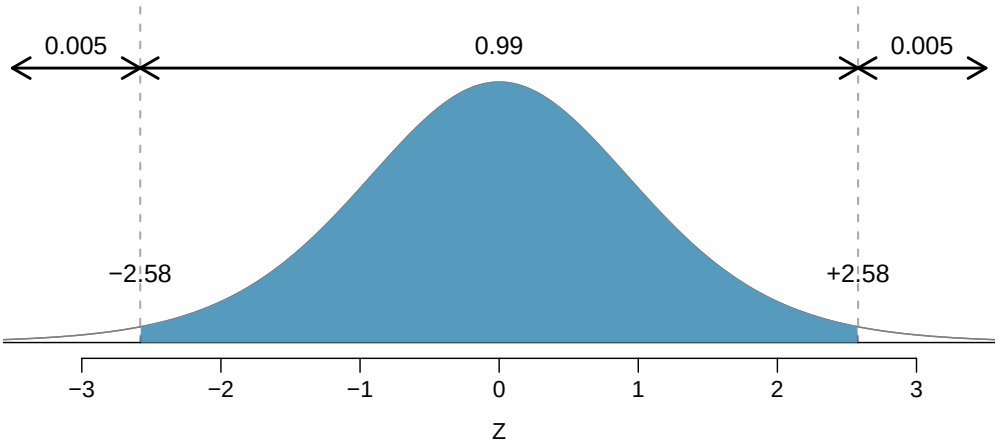


Figure 7.4: The area between $-z_{\alpha/2} = -2.58$ and $+z_{\alpha/2} = +2.58$ corresponds to a 99% confidence interval on a proportion or on the mean when σ is known

In Figure 7.5 we give a comparison of the two different levels of confidence (i.e. 95% and 99%). Notice as our level of confidence increases, the value of $|z_{\alpha/2}|$ also increases.

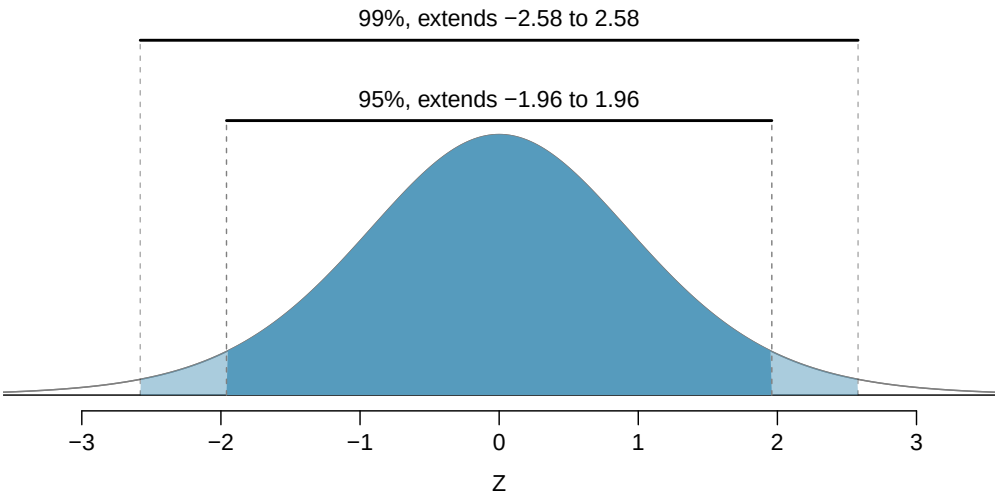


Figure 7.5: The area between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ increases as $|z_{\alpha/2}|$ becomes larger. If the confidence level is 99%, we choose $z^{\alpha/2}$ such that 99% of the normal curve is between -$z_{\alpha/2}$ and $+z_{\alpha/2}$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z_{\alpha/2} = 2.58$.

7.2 One sample confidence intervals

7.2.1 On the mean

7.2.1.1 When σ is known

In the case that we know the population standard deviation, the formula for evaluating a confidence interval is

A $(100-\alpha)\%$ confidence interval on μ when σ is known

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \tag{7.10}$$

The value of $z_{\alpha/2}$ is obtained from the standard normal tables. To find $z_{\alpha/2}$ we look at the tail probabilities of our reference distribution to determine certain cut off values. The manner in which we determine the value of $z_{\alpha/2}$ was discussed in Chapter 7.1.4. The standard error in Equation (7.10) is $\frac{\sigma}{\sqrt{n}}$ and the margin of error is $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$.

Example 7.11 Recall the Cherry 2012 Cherry Blossom Run example in the case study in Section 5.4. If the sample mean times of the 100 sample points from `run10Samp` is 95.61 minutes and the standard deviation of all the runners is 1.58 minutes, what would be an approximate 95% confidence interval for the average 10 mile time of all runners in the race? Explicitly state the standard error of the estimate, the margin of error and the 95% confidence interval

We have $n = 100$, $\bar{x} = 95.61$ and $\sigma = 1.58$. The standard error calculated using the standard deviation is

$$SE = \frac{15.78}{\sqrt{100}} = 1.58$$

From Example (7.4) we know that the value of $z_{\alpha/2} = 1.96$. Therefore the margin of error is

$$MOE = 1.96 \times \frac{15.78}{\sqrt{100}} = 3.10$$

We apply Equation ((7.10)) to get the entire confidence interval:

$$95.61 \pm 3.10 = (92.51, 98.71)$$

Based on these data, we are about 95% confident that the average 10 mile time for all runners in the race was larger than 92.45 but less than 98.77 minutes. Our interval extends out 2 standard errors from the point estimate, $\bar{x}$.

Example 7.12 Give an interpretation of the confidence interval created in Example (7.11)

In Example (7.11) we constructed a 95% confidence interval for the true mean time taken for the 10 mile run. Therefore the interpretation is that we are 95% confident that the true mean time for all runners to complete the 10 mile Cherry Blossom Run in 2012 is between 92.51 minutes and 98.71 minutes.

⊙ **Exercise 7.13** The sample data suggest the average runner's age is about 35.05 years with a standard error of 0.90 years (estimated using the sample standard deviation, 8.97). What is an approximate 95% confidence interval for the average age of all of the runners?[4]

⊙ **Exercise 7.14** Create a 99% confidence interval for the average age of all runners in the 2012 Cherry Blossom Run. The point estimate is $\bar{y} = 35.05$ and the standard error is $SE_{\bar{y}} = 0.90$.[5]

Table 7.6 gives some common values of $z_{\alpha/2}$ along with their corresponding level of confidence.

Level of Confidence	$z_{\alpha/2}$
80	1.28
90	1.65
95	1.96
99	2.58

Table 7.6: Values of $z_{\alpha/2}$ that are commonly used along with their corresponding level of confidence.

Table 7.6 is provided mainly for convenience. It is best to be able to find the value of $z_{\alpha/2}$ for any level of confidence rather than to memorize the numbers in this table. Statistics is about understanding and applying concepts rather than memorizing.

⊙ **Exercise 7.15** Use the data in Exercise (7.14) to create a 90% confidence interval for the average age of all runners in the 2012 Cherry Blossom Run.[6]

In reality there are very few situations when σ is known and the confidence interval we create for the mean is usually the one in section 7.2.1.2 when σ is unknown. However there are rare instances when we can use Equation ((7.10)) and we explored this confidence interval since we can see how the elements in Chapter 5 come together in creating a confidence interval. The confidence interval created when σ is known is also the simplest confidence interval on the mean and it is a good idea to start with something basic before moving on to more advanced situations.

FILL UP WITH SOMETHING

[4] Again apply Equation (7.1.2): $35.05 \pm 2 \times 0.90 \rightarrow (33.25, 36.85)$. We interpret this interval as follows: We are about 95% confident the average age of all participants in the 2012 Cherry Blossom Run was between 33.25 and 36.85 years.

[5] The observations are independent with $n = 100$. Apply the 99% confidence interval formula: $\bar{y} \pm 2.58 \times SE_{\bar{y}} = (32.7, 37.4)$. We are 99% confident that the average age of all runners is between 32.7 and 37.4 years.

[6] We first find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal model, $N(\mu = 0, \sigma = 1)$. We can look up -$z^\star$ in the normal probability table by looking for a lower tail of 5% (the other 5% is in the upper tail), thus $z^\star = 1.65$. The 90% confidence interval can then be computed as $\bar{y} \pm 1.65 \times SE_{\bar{y}} \rightarrow (33.6, 36.5)$. (We had already verified conditions for normality and the standard error.) That is, we are 90% confident the average age is larger than 33.6 but less than 36.5 years.

7.2.1.2 When σ is not known

In the case that we know the population standard deviation is not known, the formula for evaluating a confidence interval is

A $(100-\alpha)\%$ confidence interval on μ when σ is not known

$$\bar{x} \pm t_{(\alpha/2,\, n-1)}\left(\frac{s}{\sqrt{n}}\right) \tag{7.16}$$

Notice the similarity between Equation ((7.16)) and Equation ((7.10)). When σ is not known we use the sample standard deviation s instead and we use a value from t distribution from Chapter 4.2.3 instead of a value from the standard normal distribution. The value $t_{(\alpha/2,n-1)}$ has two subscripts. The subscript of $\frac{\alpha}{2}$ refers to the tail value of the t distribution that we are interested in and the subscript of $(n-1)$ are the degrees of freedom. This was mentioned briefly in Section 4.2.3.1 and in particular the manner in which we read the t table was discussed in Chapter 4.2.3.1. The standard error of the mean in (7.16) is $\left(\frac{s}{\sqrt{n}}\right)$ and the margin of error is $t_{(\alpha/2,\, n-1)}\left(\frac{s}{\sqrt{n}}\right)$.

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.

Since σ is not known we loose information from our sample. This is the reason that we use the t distribution instead of the standard since the extra thick tails of the t distribution are exactly the correction we need to resolve the problem of a poorly estimated standard error. An intuitive explanation to the reason we use $n-1$ degrees of freedom instead of n is because we are using s instead of σ. This means we are loosing 1 piece of information.

Figure 7.7: A Risso's dolphin.

Photo by Mike Baird (http://www.bairdphotos.com/).

Let's explore this type of confidence interval with a real world example involving marine life. Dolphins are at the top of the oceanic food chain, which causes dangerous

substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

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.[7] The data are summarized in Table 7.8. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 7.8: 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).

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 $t_{(\alpha/2,\, n-1)}$ is a cutoff we obtain based on the confidence level and the t distribution with df degrees of freedom. Before determining this cutoff, we will first need the degrees of freedom.

In our current example, we should use the t distribution with $df = 19 - 1 = 18$ degrees of freedom. Then identifying $t_{(\alpha/2,18)}$ is similar to how we found $z_{\alpha/2}$.

- For a 95% confidence interval, we want to find the cutoff $t_{(\alpha/2,18)}$ such that 95% of the t distribution is between -$t_{(\alpha/2,18)}$ and $t_{(\alpha/2,18)}$. This means we are interested in $t_{(0.025,18)}$.
- We look in the t table on page 144, find the column with area totalling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t_{(0.025,18)} = 2.10$.

Generally the value of $t_{(\alpha/2,n-1)}$ is slightly larger than what we would get under the normal model with $z_{\alpha/2}$.

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:

$$4.4 \ \pm \ 2.10 \times 0.528 = (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram. This is above the Japanese regulation level of 0.4 μg/wet gram.

⊙ **Exercise 7.17** The FDA's webpage provides some data on mercury content of fish.[8] 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

[7] Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins. Data reference: Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

[8] http://www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[9]

Example 7.18 Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Exercise (7.17). 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 $t_{(\alpha/2,\, n-1)}$.

The standard error is $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. 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 $t_{(\alpha/2,\, n-1)} = t_{(0.05,\, 14)} = 1.76$.

Example 7.19 Using the results of Exercise (7.17) and Example (7.18), compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).

Using data summaries from Exercise (7.17) and the values obtained in Example (7.18) we see that

$$\bar{x} \pm t_{0.1,\, 14} SE = 0.287 \pm 1.76 \times 0.0178 = (0.256, 0.318) \tag{7.20}$$

Example 7.21 provide an interpretation of the confidence interval calculated in Example (7.19)

We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

FILL UP WITH SOMETHING

[9] 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.

7.2.2 On a proportion

In the case that we are interested constructing a confidence interval for a proportion proportion, the formula we use is

A $(100-\alpha)\%$ confidence interval on p

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{7.22}$$

In Equation (7.22) $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the standard error of proportion and $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the margin of error. Recall that a proportion is a fraction of the overall population. We use the methods in this chapter to answer questions like the following real life example:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

Let's work out this example. According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[10] This poll included responses of 976 adults.

A sample proportion can be described as a sample mean (of sorts). If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0+1+1+s+0}{976} = 0.44$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes and failures, typically at least 10 of each. Let's now estimate the standard error of $\hat{p} = 0.44$ using Equation ((7.22)).

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$$

Let's now construct a 95% confidence interval for p, the proportion of Americans who approve of the job the Supreme Court is doing. From Chapters 7.2.1.1 and 7.1.4 we know that $z_{\alpha/2} = 1.96$ for a 95% confidence interval. Therefore the confidence interval may be computed as

$$0.44 \pm 1.96 \times 0.016 = (0.409, 0.471)$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. If the proportion has not changed since this poll, than we can say with high confidence that the job approval of the Supreme Court is below 50%.

[10] nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html

7.2.3 Assumptions

7.2.3.1 On the mean

For confidence intervals on μ the assumptions that are required are the following

1. Data is from a random sample from a large population.
2. The sample size should be less than 10% of the population.
3. Observations in the sample must be independent of each other.
4. If the sample size is small, the population distribution should be normal or at least approximately normal.

We are mainly concerned on sample size in assumption 4 because the effect of the central limit theorem will provide us with a sampling distribution of $\bar{x}$ that is approximately normal for sufficiently large n. When σ is not known and we using the t distribution is still valid because the t begins to resemble the standard normal distribution.

7.2.3.2 On a proportion

For confidence intervals on p the assumptions that are required are the following

1. Data is from a random sample from a large population.
2. Observations in the sample are independent of each other.
3. $np \geq 10$ and $n(1-p) \geq 10$

Since p is unknown we can attempt to verify assumption 3 by estimating p with $\hat{p}$ and checking whether $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

7.3 Two sample confidence intervals

7.3.1 On a 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 (i.e. our parameter of interest is $\mu_1 - \mu_2$). Paired data are discussed in Chapter 7.3.2. The methods are similar in theory to those of Chapter 7.2 but different in the details.

7.3.1.1 When σ_1 and σ_2 are known

This is the most basic two sample confidence interval on a difference of means.

A $(100-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ when σ_1 and σ_2 are known

$$\bar{x}_1 - \bar{x}_2 \quad \pm \quad z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.23}$$

In Equation ((7.23)) the term under the square root is the standard error on the difference of means and this term multiplied by $z_{\alpha/2}$ is the margin of error. There are very few cases where Equation ((7.23)) is used since it is very rare to know the standard deviations of two populations.

7.3.1.2 When σ_1 and σ_2 are not known

When are interested in the difference of means of two populations and the standard deviations of both populations are not known we have to consider two possibilities. They are when $\sigma_1 \neq \sigma_2$ or when $\sigma_1 = \sigma_2$. The case in which the standard deviations are both not known are broken down into 2 further sub-cases in which the standard deviations are known to be (or at least assumed) equal or not equal. Since this is an introductory text in statistics, we may consider that we either know or at least we can assume $\sigma_1 \neq \sigma_2$ or $\sigma_1 = \sigma_2$ based on additional information available from the study or expert knowledge in the area of interest. Typically when there is knowledge of equal population standard deviations, we usually observe sample standard deviations that are also relatively close.

Caution about assuming equal standard deviations
In general do not assume $\sigma_1 = \sigma_2$.

If we do not have any information about equal or unequal population standard deviations we should not make assumptions about them. There is a *crude rule of thumb* that can be applied which is that if If the larger sample standard deviation divided by the smaller sample standard deviation is greater or equal to two then we should not assume $\sigma_1 = \sigma_2$. By this we mean that if

$$\frac{\max(s_1,\, s_2)}{\min(s_1,\, s_2)} \geq 2 \quad \rightarrow \quad \text{do not assume } \sigma_1 = \sigma_2$$

This rule of thumb states that if our sample standard deviations are different enough we can not safely assume $\sigma_1 = \sigma_2$. There are statistical tests available to assess the assumption of equality of variance however such tests are more suited for more advanced texts in statistics.

7.3.1.2.1 When $\sigma_1 \neq \sigma_2$

A $(100 - \alpha)\%$ confidence interval on $\mu_1 - \mu_2$ when σ_1 and σ_2 are not known and $\sigma_1 \neq \sigma_2$

$$\bar{x}_1 - \bar{x}_2 \quad \pm \quad t_{(\alpha/2,\, d)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{7.24}$$

where a conservative estimate of d is

$$d = \min(n_1 - 1,\, n_2 - 1) \tag{7.25}$$

Note that we stated that d is a conservative estimate of the degrees of freedom. A more accurate value of d is given by

$$d = \frac{\left(\frac{s_1^2}{x_1} + \frac{s_2^2}{x_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \tag{7.26}$$

The value calculated be Equation ((7.26)) is typically a real number and not an integer and this is perfectly fine (Recall Chapter 4.2.3). We can either be conservative and take the floor[11] of the value calculated in Equation ((7.26)) or we can use software to obtain a more accurate value.

Caution: Welch's Interval and the Behrens—Fisher problem
This method outlined in Equation (7.24) is known as **WelchÕs Interval** and it is a good approximation for the situation in this chapter. In reality interval estimation (as well as hypothesis testing) on a difference of two means from two independent samples from two normally distributed populations when $\sigma_1 \neq \sigma 2$ is an open problem called the **BehrensFisher problem**.

7.3.1.2.2 When $\sigma_1 = \sigma_2$

When we either know that the two population standard deviations are equal or when we can assume that they are equal, the formula for the confidence interval is

A $(100 - \alpha)\%$ confidence interval on $\mu_1 - \mu_2$ when σ_1 and σ_2 are not known and $\sigma_1 = \sigma_2$

$$\bar{x}_1 - \bar{x}_2 \quad \pm \quad t_{(\alpha/2,\, n_1+n_2-2)} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{7.27}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{7.28}$$

The value s_p^2 is called the pooled sample variance. It is an average of s_1 and s_2 that takes the sample sizes into account. Although it may appear intimidating it is actually a matter of substituting the correct values and performing calculations.

We will apply Equation ((7.24)) to construct a 95% confidence interval on a difference of means to the participants in the 2012 Cherry Blossom Run from the case study in Chapter 5.4. We would like to estimate the average difference in run times for men and women using the `run10Samp` data set, which was a simple random sample of 45 men and 55 women from all runners in the 2012 Cherry Blossom Run. Table 7.9 presents relevant summary statistics, and box plots of each sample are shown in Figure 7.10.

[11] The floor refers to rounding down to the nearest integer.

	men	women
$\bar{x}$	87.65	102.13
s	12.5	15.2
n	45	55

Table 7.9: Summary statistics for the run time of 100 participants in the 2009 Cherry Blossom Run.

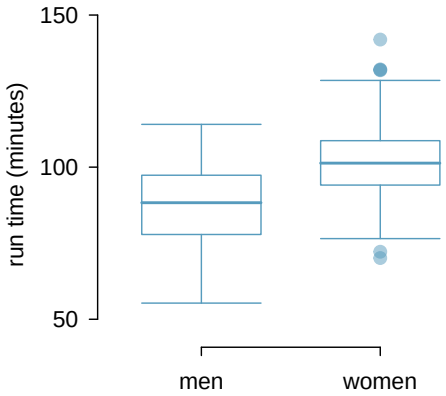


Figure 7.10: Side-by-side box plots for the sample of 2009 Cherry Blossom Run participants.

The two samples are independent of one-another, so the data are not paired. Instead a point estimate of the difference in average 10 mile times for men and women, $\mu_w - \mu_m$, can be found using the two sample means:

$$\bar{x}_w - \bar{x}_m \;=\; 102.13 - 87.65 \;=\; 14.48$$

Suppose we are told by an expert in sports science that the standard deviations of the time taken to complete 10 mile marathons is approximately equal. We can also visually see that $s_w \approx s_m$. As such we need to calculate the pooled sample variance. Using Equation ((7.28)) we get

$$\begin{aligned} s_p^2 &= \frac{(n_w - 1)s_w^2 + (n_m - 1)s_m^2}{n_m + n_m - 2} \\ &= \frac{(55 - 1)15.2^2 + (45 - 1)12.5^2}{45 + 55 - 2} \\ &= 197.46 \end{aligned}$$

We can quantify the variability in the point estimate, $\bar{x}_w - \bar{x}_m$, using the following formula for its standard error:

$$\begin{aligned} SE_{\bar{x}_w - \bar{x}_m} &= \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &= \sqrt{197.46\left(\frac{1}{55} + \frac{1}{45}\right)} \\ &= 2.82 \end{aligned}$$

Here a point estimate, $\bar{x}_w - \bar{x}_m = 14.48$, is associated with standard error $SE = 2.82$. The next step is to find the appropriate value from the t tables. For a 95% confidence interval, $\alpha/2 = 0.025$. The appropriate value of the degrees of freedom d that we should use is

$$n_w + n_m - 2 = 55 + 45 - 2 = 98$$

This means that the value we use from the t tables is $t_{0.025,98}$. However Table A.3 does not contain values for 98 degrees of freedom. Therefore we round down to the nearest available degrees of freedom in the table which are 90 degrees of freedom. The value of $t_{(0.025,90)} = 1.99$. Our 95% confidence interval is therefore

$$14.48 \ \pm \ 1.99 \times 2.82 = (8.87,\ 20.09)$$

Based on the samples, we are 95% confident that men ran, on average, between 8.87 and 20.09 minutes faster than women in the 2012 Cherry Blossom Run.

7.3.2 On paired data

A confidence interval on **paired data** (also referred to as a termmatched pairs) is performed when we have two measurements which depend on each of the units sampled. Our population is the population of differences and our parameter of interest is the mean difference between μ_d.

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.

With paired data we take the difference in the two measurements for each unit. This gives us a set of differences. We then calculate the sample mean and sample standard deviation for these differences which we label as $\bar{x}_d$ and s_d respectively[12]. We then use $\bar{x}_d$ and s_d to create a confidence interval on the mean difference.

A $(100 - \alpha)\%$ confidence interval on paired data

$$\bar{x}_d \ \pm \ t_{(\alpha/2,\, n-1)}\left(\frac{s_d}{\sqrt{n}}\right) \tag{7.29}$$

[12] We use a subscript of d to indicate that we our statistics are on a set of differences. The calculations of the mean and standard deviation are still the same.

Notice that Equation (7.29) is essentially the equation for constructing a confidence interval on the mean when σ is not known which is given by Equation (7.16). We have however introduced different labels that fit this context.

The example we will use to illustrate Equation ((7.29)) is whether textbooks are actually cheaper online? We compare the price of textbooks at the bookstore of the University of California, Los Angeles (UCLA) and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses.[13] A portion of this data set is shown in Table 7.11.

	Department	Course	UCLA	Amazon.com	Difference
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

Table 7.11: Six entries from the `textbooks` data set.

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. Since these two sets of observations have this special correspondence, they can be considered as paired.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the `textbook` data set, we look at the difference in prices, which is represented as the `diff` variable in the `textbooks` data. Here the differences are taken as

$$\text{UCLA price} - \text{Amazon price}$$

for each book. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure 7.12. Using differences between paired observations is a common and useful way to analyze paired data.

The distribution of differences, shown in Figure 7.12, is slighlty skewed, but this amount of skew is reasonable for this sized data set ($n = 73$).

[13] When a class had multiple books, only the most expensive text was considered.

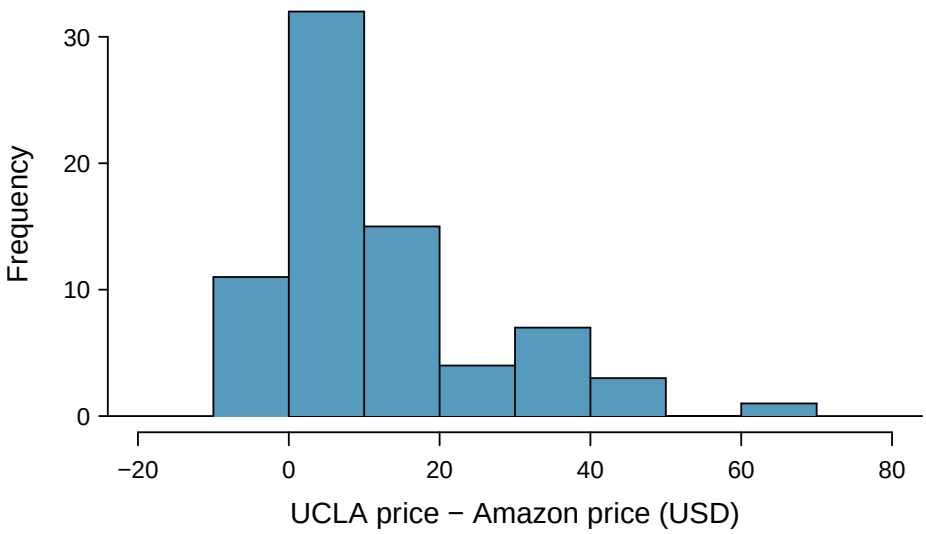


Figure 7.12: Histogram of the difference in price for each book sampled. These data are slighlty skewed.

A summary of the column of differences in Table 7.11 is given in Table 7.13 below.

n_d	$\bar{x}_d$	s_d
73	12.76	14.26

Table 7.13: Summary statistics for the price differences. There were 73 books, so there are 73 differences.

We compute the standard error associated with $\bar{x}_d$ using the standard deviation of the differences ($s_d = 14.26$) and the number of differences ($n = 73$):

$$SE_{\bar{x}_d} = \frac{s_d}{\sqrt{n}} = \frac{14.26}{\sqrt{73}} = 1.67$$

Let's now create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon. We need the value of $t_{\alpha/2,n-1} = t_{0.025,72}$. However our the t distribution table in Chapter A.3 does not provide values for 72 degrees of freedom. Therefore we round down to 70 degrees of freedom and use $t_{\alpha/2,n-1} = t_{0.025,72} = 1.99$. Our 95% confidence interval on the true difference between the bookstore and Amazon.com is

$$12.76 \ \pm \ 1.99 \times 1.67 = (9.44, 16.08)$$

We are 95% confident that Amazon is, on average, between \$9.49 and \$16.03 cheaper than the UCLA bookstore for UCLA course books.

FILL UP

7.3.3 On a difference of two proportions

In this Chapter we analyze the method to make conclusions about the difference in two population proportions: $p_1 - p_2$.

We examine an example in which we compare the approval of the 2010 healthcare law under two different question phrasings.

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.

The formula for constructing a confidence interval on a difference of proportions is

A $(100 - \alpha)\%$ confidence interval on $p_1 - p_2$ when σ_1 and σ_2 are not known and $\sigma_1 = \sigma_2$

$$\hat{p}_1 - \hat{p}_2 \quad \pm \quad z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{7.30}$$

In Equation ((7.30)), everything under the square root sign is the standard error of proportion and this value multiplied by $z_{\alpha/2}$ is the margin of error.

Let's now do an example. In the setting of confidence intervals, the sample proportions are used to verify the success-failure condition and also compute standard error, just as was the case with a single proportion. The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[14]

> As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed. Table 7.14 shows the results of this experiment. We shall create and interpret a 90% confidence interval of the difference in approval.

	Sample size (n_i)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Table 7.14: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

[14] www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/. Sample sizes for each polling group are approximate.

We will let p_1 corresponds to the original ordering and p_2 to the reversed ordering. A point estimate for the difference in proportions is

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13$$

The standard error may be computed from Equation ((7.30))using the sample proportions:

$$SE = \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025$$

For a 90% confidence interval, $z_{\alpha/2} = 1.65$. Therefore our 90% confidence interval on the difference of proportions is:

$$0.13 \ \pm \ 1.65 \times 0.025 = (0.09, 0.17)$$

The interpretation of this interval is that we are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still fluid on the health insurance mandate.

7.3.4 Assumptions

7.3.4.1 On a difference of two means

For confidence intervals on $\mu_1 - \mu_2$ the assumptions that are required are the following

1. Data from both samples are taken from random samples from large populations.
2. The sample size of each sample should be less than 10% of their corresponding population.
3. Observations in a sample must be independent of each observation from the same sample.
4. Observations in a sample must be independent of each observation from the other sample.
5. Both populations are approximately normally distributed.

7.3.4.2 On paired data

For confidence intervals on μ_d the assumptions that are required are the following

1. Two measurements from each observation are dependent on the unit from which they were measured.
2. The sample size should be less than 10% of the population.
3. Measurements on each unit are independent of each measurement on other units.
4. The population of differences is normally distributed.
5. For the Pooled Method : The two populations have the same variance.[15]
   For Welch's Interval : The two populations do not have the same variance.[16]

[15] This assumption is called the assumption of *homogeneity* of variance
[16] This assumption is called the assumption of *heterogeneity* of variance

7.3.4.3 On a difference of two proportions

For confidence intervals on $p_1 - p_2$ the assumptions that are required are the following

1. Data from both samples are taken from random samples from large populations.
2. Observations in a sample must be independent of each observation from the same sample.
3. Observations in a sample must be independent of each observation from the other sample.
4. $n_1 p_1 \geq 10$ and $n_1(1 - p_1) \geq 10$ as well as $n_2 p_2 \geq 10$ and $n_2(1 - p_2) \geq 10$.

Since we do not know p_1 or p_2 we can attempt to verify assumption 4 by estimating p_1 with $\hat{p}_1$ as well as estimating p_2 with $\hat{p}_2$. As such we can check whether $n_1\hat{p}_1 \geq 10$ and $n_1(1-\hat{p}_1) \geq 10$ as well $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$ (Similar to the check we can perform in assumption 3 in Chapter 7.2.3.2)

Chapter 8

Hypothesis testing

8.1 Introduction

In this Chapter we use statistics to make conclusions on parameters. In Chapter 7 we constructed a range of values which we stated captured the mean with a certain level of confidence. In this chapter we will learn techniques to quantify the strength of a conclusion using probability.

The mathematical background behind a hypothesis test starts with a decision rule which is something that is used to decide on a conclusion to make regarding the value of a parameter based on the result of a test. A *test statistic* is a value calculated using sample data as well as a value from a status quo belief. Test statists are derived using a decision rule as well as something called the likelihood ratio test. These topics are more suited for advanced undergraduate of graduate courses in mathematical statistics but me mention them so that the reader is aware about the rigorous mathematical framework behind hypothesis tests.

For an introductory course we will skip the mathematical derivation of a hypothesis test and get right into the procedure of conducting them. This makes the process of hypothesis testing somewhat algorithmic to some extent. There are several steps involved in performing a hypothesis test. We will list each of the steps involved and discuss each of them in more detail.

1. State the null and alternative hypothesis
2. Find the appropriate test statistic
3. Find the p-value associated with the hypothesis test
4. Compare the p-value to a level of significance (α)
5. Make a conclusion

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 favour of the alternative. The hallmarks of hypothesis testing are also found in the US court system. In Chapters ??? we will learn about hypothesis tests on different parameters however the procedure involved is the same.

8.1.1 State the null and alternative hypothesis

The first step in hypothesis testing is to state the null and alternative hypothesis. The **null hypothesis** is the conservative or skeptical belief and the **alternative hypothesis** is the claimed belief. The alternative is usually a researchers' belief. The null hypothesis is represented by H_0 and the alternative hypothesis is represented by H_a.[1]

Null and alternative hypotheses

H_0 : Null hypothesis
Represents either a skeptical or conservative perspective on a claim to be tested.

H_a : Alternative hypothesis
Represents an alternative claim under consideration and is often represented by a range of possible parameter values.

When we stated that the null null hypothesis represents the conservative belief; by this we mean that it represents the safe belief. For example, during the development of a pharmaceutical drug the conservative belief is to believe that the drug has no affect and does not work.

In the hypothesis testing framework we start by assuming that H_0 is true (since it is the safe belief) and then we continue to find evidence either supporting H_0 or against H_0.

Suppose γ is the parameter we are interested in and γ_0 is the hypothesized value of γ under the null hypothesis. We are being very general here and γ can represent the population meanμ or the population proportion p is a difference of population means $\mu_1 - mu_2$ etc. γ_0 is some numeric value γ is supposed to equal under the null hypothesis. There are 3 possible hypothesis tests we can perform on γ and these are:

1. $H_0 = \gamma_0$ vs. $H_a > \gamma_0$
2. $H_0 = \gamma_0$ vs. $H_a < \gamma_0$
3. $H_0 = \gamma_0$ vs. $H_a \neq \gamma_0$

We can only perform one of the hypothesis tests above at a time (i.e. we do not conduct all 3 hypothesis tests simultaneously). Hypothesis tests 1 and 2 are called **one tailed tests** or **one sided tests** and hypothesis test 3 is called the **two tailed test** or **two sided test**. These terms will become more familiar in Chapter ???.

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

The null hypothesis often represents a skeptical position or a perspective of no difference. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change.

[1] Some textbooks might represent the alternative hypothesis by H_1 rather than H_a. This is also acceptable standard notation. However we will use H_a to represent the alternative hypothesis.

8.1.2 Find the appropriate test statistic

The test statistic that we calculate depends on the situation and the information available. An estimator and the assumed parameter value under the null hypothesis is used in the calculation of the test statistic.

Basic skeleton of a test statistic

$$\text{test statistic} = \frac{(\text{an estimator}) - \begin{pmatrix}\text{hypothesized value of the parameter} \\ \text{under the null hypothesis}\end{pmatrix}}{(\text{Standard error of estimate})} \tag{8.1}$$

Compare (8.1) with the basic skeleton of a confidence interval (i.e. (7.2) on page 182) and notice the similarities. hypothesized value of the parameter under the null hypothesis refers to the numeric value of γ_0 that was discussed in Chapter 8.1.1. So two things that the test statistic depend on are the data that is used to calculate the estimator and γ_0.

The calculated test statistic follows a certain reference distribution.

Summary of reference distributions to use for hypothesis tests

When σ is known	$\longrightarrow$	estimator $\sim$ Z distribution
When σ is not known	$\longrightarrow$	estimator $\sim$ t distribution
For proportions	$\longrightarrow$	estimator $\sim$ Z distribution

We use the test statistic along with its reference distribution to calculate the p-value which is something we discuss in Chapter 8.1.3.

8.1.3 Find the p-value

The **p-value** is a way of quantifying the strength of the evidence against the null hypothesis and in favour of the alternative. The p-value depends on the the test statistic calculated (see Chapter 8.1.2) and the alternative hypothesis (see Chapter 8.1.1). Formally the *p-value* is a conditional probability.

p-value

The p-value is the probability of observing a test statistic at least as extreme as the one calculated with the sample data collected purely by chance under the assumption that H_0 is true.

It is important to note that we used the words "at least as extreme" and not just "as extreme" in the definition of the p-value in the text box above. An alternate way to defining the p-value is that it is the probability of observing data at least as favourable to the alternative hypothesis as our current data set, if the null hypothesis is true. The estimator mentioned in Chapter 8.1.2 is used to be representative of the data.

Also recall from Chapter 8.1.2 that the test static calculated with (8.1) follows a certain distribution. The test statistic calculated for the hypothesis test is used along with

the alternative hypothesis to determine the p-value. We illustrate this in Figure **??** below. Note that we are not being specific about the distribution used and the curve just represents any valid reference distribution for the test statistic calculated. We use a bell-shaped curve in this figure because the only reference distributions that we cover in this text are the standard normal distribution and the t distribution and both of these are bell-shaped.

If we are conducting a one tailed test, the p-value is the area to the right of the test statistic on the reference distribution if the alternative is that the parameter is greater than the hypothesized value. This is illustrated in Figures 8.1 and 8.2 below.

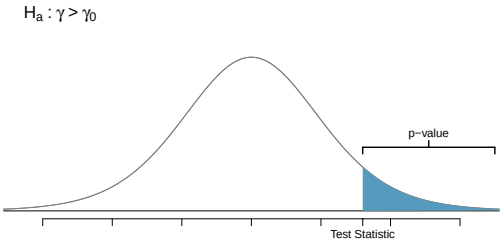


Figure 8.1: Finding the p-value when $H_a : \gamma > \gamma_0$

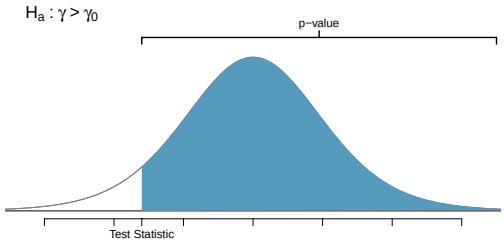


Figure 8.2: Finding the p-value when $H_a : \gamma > \gamma_0$ and we have a large tail

The p-value will be the area to the left of the test statistic on the reference distribution if the alternative is that the parameter is less than the hypothesized value. This is illustrated in Figures 8.3 and 8.4 below.

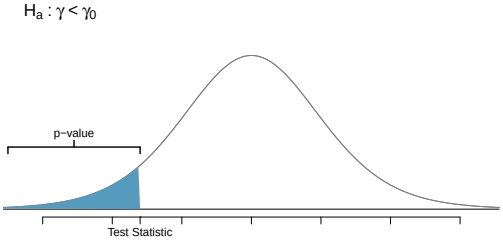


Figure 8.3: Finding the p-value when $H_a : \gamma < \gamma_0$

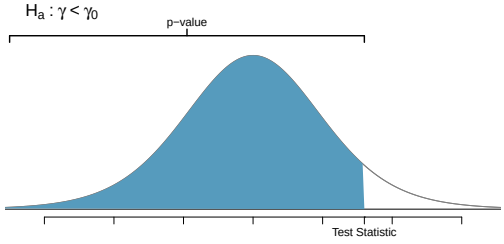


Figure 8.4: Finding the p-value when $H_a : \gamma < \gamma_0$ and we have a large tail

Caution: Finding the p-value for one tailed tests

For one tailed tests don't automatically assume that the p-value is the area in the smaller tail. Refer back to the alternative hypothesis to decide on the area that will give the p-value.

When we conduct a two tailed we consider the most extreme cases possible. This means we look at the area in the tail to the right of |test statistic| as well as the area in the tail to the left of -|test statistic|. This is illustrated in Figure 8.5.

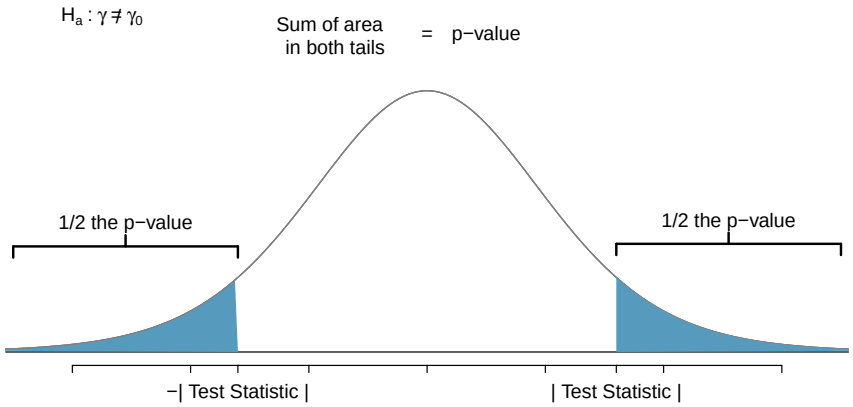


Figure 8.5: Finding the p-value when $H_a : \gamma \neq \gamma_0$

To find the p-value for a twi tailed test we have to add up the area in both tails. For symmetric distributions such as the normal and t distribution we can find the area of one tail and multiply that result by two.

TIP: Two tailed tests
When we conduct a two tailed test, make sure that out picture has two tails.

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

8.1.4 Compare with a level of significance α

The level of significance is a cut off value (i.e. tolerance) that is used to decide whether we should reject the null hypothesis or not. The level of significance is associated with the symbol α. The p-value is compared with α to determine whether to reject the null hypothesis or not. The value of α is usually predetermined before conducting a test. A commonly used value is $\alpha = 0.05$ (which corresponds to a fraction of $1/20$). [2]

A natural question to ask is whether we can declare a level of significance for a test after we calculate the p-value. The short answer is that we can, however this is typically not done and it is not considered a good practice in general since it may give the appearance that results are being doctored to favour a researchers' beliefs.

[2]The is the reason we often see the common "19 times out of 20" text in articles.

8.1.5 Make a conclusion

The conclusion we make on our hypothesis test depends on our p-value and α. A summary of the manner in which confusions are made is given below.

Making a conclusion on a hypothesis test

p-value $> \alpha$ $\longrightarrow$ The evidence supports the null hypothesis.
Do not reject the null hypothesis.

p-value $> \alpha$ $\longrightarrow$ The evidence is against the null hypothesis.
The evidence rejects the null hypothesis.

Notice how we use conservative language and state that we do "not reject the null hypothesis" or the "evidence is against the null hypothesis". It is incorrect to state that the "null hypothesis is wrong" or "The null hypothesis is right". This kind of language indicates that we are completely certain about the result of a test. We are working with probabilities so we do not state that our conclusion is completely correct for certain since we may have obtained a bad sample that is not representative of the data resulting in an incorrect conclusion. We will learn more about the errors that could occur in a hypothesis test in Chapter ???.

p-value as a tool in hypothesis testing

The p-value quantifies how strongly the data favour H_a over H_0. The smaller the p-value the stronger the evidence against H_0.

A sufficiently small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_a.

Another reason that we say we do not "reject the null hypothesis" is because the evidence we have does not indicate that we definitely have the correct null hypothesis; it's more to do with the evidence is not at a sufficiently strong level to be against it. This is similar to a jury making a decision in court. A person is considered innocent until proven guilty, so a person is considered innocent under the null. Evidence presented may be circumstantial, however if this evidence may not be strong enough to raise enough doubt beyond a reasonable level, the jury should consider the person to be innocent.

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.

TIP: Hypothesis testing framework

The skeptic will not reject the null hypothesis (H_0), unless the evidence in favor of the alternative hypothesis (H_a) is so strong that she rejects H_0 in favor of H_a.

Although we make this comparison, a statistical test of hypothesis is usually less subjective than circumstantial evidence in a trial.

8.1.6 Testing hypotheses using confidence intervals

A simple way to conduct a two-sided hypothesis test is by using a confidence interval. To state briefly, the conclusion we make is that the evidence would fail to reject the null hypothesis of equality at a significance level of α if the hypothesized value of the parameter under the null hypothesis is inside an approximate $(100 - \alpha)\%$ confidence interval for the parameter (and vice versa if the hypothesized value of the parameter under the null hypothesis is not inside the interval).

We will illustrate this procedure with an example. Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Run, comparing runners in 2006 and 2012. Technological advances in shoes, training, and diet might suggest runners would be faster in 2012. An opposing viewpoint might say that with the average body mass index on the rise, people tend to run slower. In fact, all of these components might be influencing run time.

In addition to considering run times in this section, we consider a topic near and dear to most students: sleep. A recent study found that college students average about 7 hours of sleep per night.[3] However, researchers at a rural college are interested in showing that their students sleep longer than seven hours on average.

The average time for all runners who finished the Cherry Blossom Run in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine if the `run10Samp` data set provides strong evidence that the participants in 2012 were faster or slower than those runners in 2006, versus the other possibility that there has been no change.[4] We simplify these three options into two competing **hypotheses**:

H_0: The average 10 mile run time was the same for 2006 and 2012.

H_a: The average 10 mile run time for 2012 was *different* than that of 2006.

⊙ **Exercise 8.2** 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?[5]

In the example with the Cherry Blossom Run, the null hypothesis represents no difference in the average time from 2006 to 2012. The alternative hypothesis represents something new or more interesting: there was a difference, either an increase or a decrease. These hypotheses can be described in mathematical notation using μ_{12} as the average run time for 2012:

H_0: $\mu_{12} = 93.29$

H_a: $\mu_{12} \neq 93.29$

where 93.29 minutes (93 minutes and about 17 seconds) is the average 10 mile time for all runners in the 2006 Cherry Blossom Run. Using this mathematical notation, the hypotheses

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

[4] While we could answer this question by examining the entire population data (`run10`), we only consider the sample data (`run10Samp`), which is more realistic since we rarely have access to population data.

[5] 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).

can now be evaluated using statistical tools. We call 93.29 the **null value** since it represents the value of the parameter if the null hypothesis is true. We will use the `run10Samp` data set to evaluate the hypothesis test.

We can start the evaluation of the hypothesis setup by comparing 2006 and 2012 run times using a point estimate from the 2012 sample: $\bar{x}_{12} = 95.61$ minutes. This estimate suggests the average time is actually longer than the 2006 time, 93.29 minutes. However, to evaluate whether this provides strong evidence that there has been a change, we must consider the uncertainty associated with $\bar{x}_{12}$.

We learned in Section 6.1 that there is fluctuation from one sample to another, and it is very unlikely that the sample mean will be exactly equal to our parameter; we should not expect $\bar{x}_{12}$ to exactly equal μ_{12}. Given that $\bar{x}_{12} = 95.61$, it might still be possible that the population average in 2012 has remained unchanged from 2006. The difference between $\bar{x}_{12}$ and 93.29 could be due to *sampling variation*, i.e. the variability associated with the point estimate when we take a random sample.

In Section 7, confidence intervals were introduced as a way to find a range of plausible values for the population mean. In Example (7.11) using data in `run10Samp`, a 95% confidence interval for the 2012 population mean, μ_{12}, was calculated as

$$(92.51, 98.71)$$

Recall that in this example, we knew that the value of σ is 1.58 minutes. Since the 2006 mean of 93.29 falls in the range of plausible values, we cannot say the null hypothesis is implausible. That is, we failed to reject the null hypothesis, H_0.

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

Example 8.3 Next consider whether there is strong evidence that the average age of runners has changed from 2006 to 2012 in the Cherry Blossom Run. In 2006, the average age was 36.13 years, and in the 2012 `run10Samp` data set, the average was 35.05 years for 100 runners. Suppose that the standard deviation of all runners is 8.97 years.

First, set up the hypotheses:

H_0: The average age of runners has not changed from 2006 to 2012, $\mu_{age} = 36.13$.
H_a: The average age of runners has changed from 2006 to 2012, $\mu_{age} \neq 36.13$.

We have previously verified conditions for this data set. The normal model may be applied to $\bar{y}$ and the estimate of SE should be very accurate. Using the sample mean and standard error, we can construct a 95% confidence interval for μ_{age} to determine if there is sufficient evidence to reject H_0:

$$\bar{y} \pm 1.96 \times \frac{\sigma}{\sqrt{100}} = 35.05 \pm 1.96 \times 0.90 = (33.29, 36.81)$$

This confidence interval contains the *null value*, 36.13. Because 36.13 is not implausible, we cannot reject the null hypothesis. We have not found strong evidence that the average age is different than 36.13 years.

⊙ **Exercise 8.4** Colleges frequently provide estimates of student expenses such as housing. A consultant hired by a community college claimed that the average student housing expense was $650 per month. What are the null and alternative hypotheses to test whether this claim is accurate?[6]

⊙ **Exercise 8.5** The community college decides to collect data to evaluate the $650 per month claim. They take a random sample of 75 students at their school and obtain the data represented in Figure 8.6. Can we apply the normal model to the sample mean?[7]

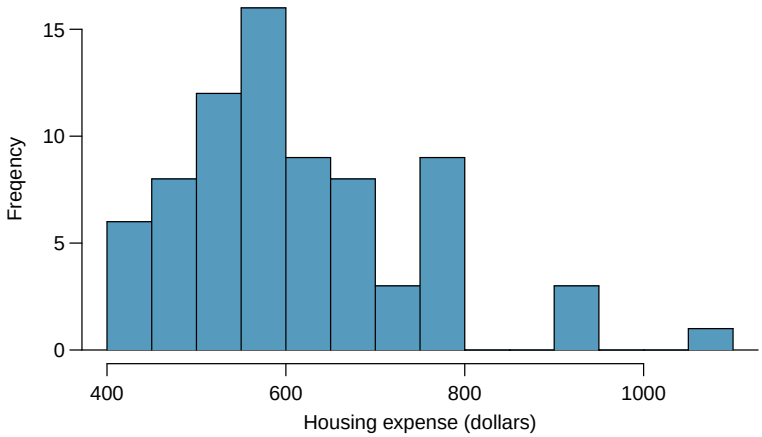


Figure 8.6: Sample distribution of student housing expense. These data are moderately skewed, roughly determined using the outliers on the right.

● **Example 8.6** The sample mean for student housing is $611.63 and the population standard deviation is $132.85. Construct a 95% confidence interval for the population mean and evaluate the hypotheses of Exercise (8.4).

The standard error associated with the mean may be estimated using the sample standard deviation divided by the square root of the sample size. Recall that $n = 75$ students were sampled.

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{132.85}{\sqrt{75}} = 15.34$$

You showed in Exercise (8.5) that the normal model may be applied to the sample mean. This ensures a 95% confidence interval may be accurately constructed:

$$\bar{x} \pm z_{\alpha/2} SE = 611.63 \pm 1.96 \times 15.34 = (581.56, 641.70)$$

Because the null value $650 is not in the confidence interval, a true mean of $650 is implausible and we reject the null hypothesis. The data provide statistically significant evidence that the actual average housing expense is less than $650 per month.

[6] H_0: The average cost is $650 per month, $\mu = \$650$.
H_a: The average cost is different than $650 per month, $\mu \neq \$650$.

[7] Applying the normal model requires that certain conditions are met. Because the data are a simple random sample and the sample (presumably) represents no more than 10% of all students at the college, the observations are independent. The sample size is also sufficiently large ($n = 75$) and the data exhibit only moderate skew. Thus, the normal model may be applied to the sample mean.

8.2 One sample hypothesis tests

8.2.1 On the mean

8.2.1.1 When σ is known

In the case that we know the population standard deviation, the test statistic that is used to perform hypothesis tests on the mean is:

Test statistic for a hypothesis test on μ when σ is known

$$Z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{8.7}$$

Reference distribution: Standard normal distribution

Let's do an example to illustrate all the steps we listed in Chapter 8.1. We will work out an example related to

8.2.1.2 When σ is not known

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

A skeptic would have no reason to believe that sleep patterns at this school are different than the sleep patterns at another school. We can set up the null hypothesis for this test as a skeptical perspective: the students at this school average 7 hours of sleep per night. The alternative hypothesis takes a new form reflecting the interests of the research: the students average more than 7 hours of sleep. We can write these hypotheses as

H_0: $\mu = 7$.

H_a: $\mu > 7$.

Since our alternative is $\mu > 7$, we are conduction a one tailed hypothesis test. In this investigation, there is no apparent interest in learning whether the mean is less than 7 hours.[8]

The researchers at the rural school conducted a simple random sample of $n = 110$ students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours. A histogram of the sample is shown in Figure 8.7.

Before we can use a normal model for the sample mean or compute the standard error of the sample mean, we must verify conditions. (1) Because this is a simple random sample from less than 10% of the student body, the observations are independent. (2) The sample size in the sleep study is sufficiently large since it is greater than 30. (3) The data show moderate skew in Figure 8.7 and the presence of a couple of outliers. This skew and the outliers (which are not too extreme) are acceptable for a sample size of $n = 110$. With

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

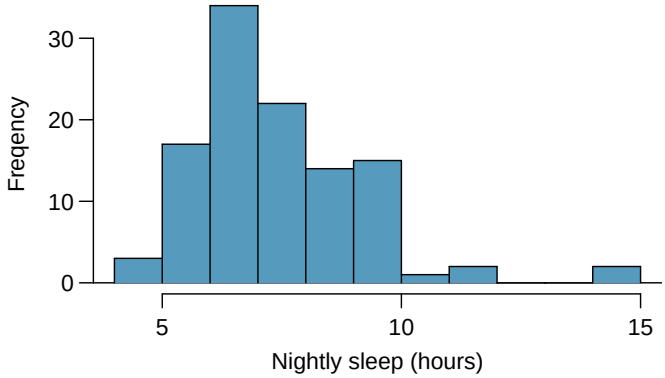


Figure 8.7: Distribution of a night of sleep for 110 college students. These data are moderately skewed.

these conditions verified, the normal model can be safely applied to $\bar{x}$ and the estimated standard error will be very accurate.

⊙ **Exercise 8.8** What is the standard deviation associated with $\bar{x}$? That is, estimate the standard error of $\bar{x}$.[9]

The hypothesis test will be evaluated using a significance level of $\alpha = 0.05$. We want to consider the data under the scenario that the null hypothesis is true. In this case, the sample mean is from a distribution that is nearly normal and has mean 7 and standard deviation of about 0.17. Such a distribution is shown in Figure 8.8.

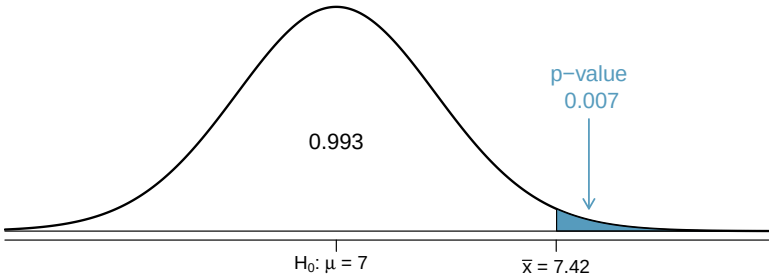


Figure 8.8: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

The shaded tail in Figure 8.8 represents the chance of observing such a large mean, conditional on the null hypothesis being true. That is, the shaded tail represents the p-

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

value. We shade all means larger than our sample mean, $\bar{x} = 7.42$, because they are more favorable to the alternative hypothesis than the observed mean.

We compute the p-value by finding the tail area of this normal distribution, which we learned to do in Section 4.2.2. First compute the Z score of the sample mean, $\bar{x} = 7.42$:

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

Using the normal probability table, the lower unshaded area is found to be 0.993. Thus the shaded area is $1 - 0.993 = 0.007$. *If the null hypothesis is true, the probability of observing such a large sample mean for a sample of 110 students is only 0.007.* That is, if the null hypothesis is true, we would not often see such a large mean.

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

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

The ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors H_a.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative.
- We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.
- We should always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether or not to reject H_0. This method ensures that the Type 1 Error rate does not exceed the significance level standard.

8.3 Decision errors

Hypothesis tests are not flawless. Just think of the court system: innocent people are sometimes wrongly convicted and the guilty sometimes walk free. Similarly, we can make a wrong decision in statistical hypothesis tests. However, the difference is that we have the tools necessary to quantify how often we make such errors.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios in a hypothesis test, which are summarized in Table 8.9.

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

Table 8.9: 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.

⊙ **Exercise 8.9** In a US court, the defendant is either innocent (H_0) or guilty (H_a). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Table 8.9 may be useful.[10]

⊙ **Exercise 8.10** How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?[11]

⊙ **Exercise 8.11** How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[12]

Exercises (8.9)-(8.11) 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. 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 8.4.

α
significance level of a hypothesis test

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

[11] 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.

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

A confidence interval is, in one sense, simplistic in the world of hypothesis tests. Consider the following two scenarios:

- The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow say, quantitatively, that it was a close decision.

- The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close. Such a case is depicted in Figure 8.10.

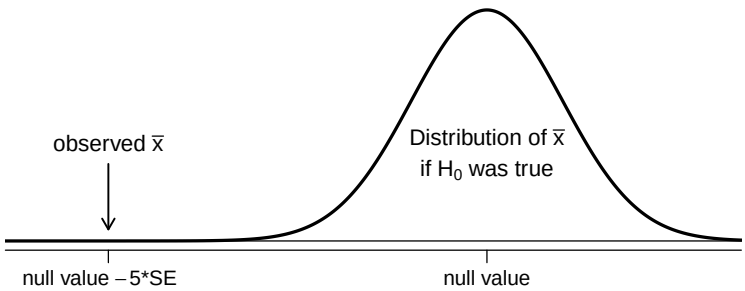


Figure 8.10: It would be helpful to quantify the strength of the evidence against the null hypothesis. In this case, the evidence is extremely strong. This figure is illustrative of the reason we use p-values.

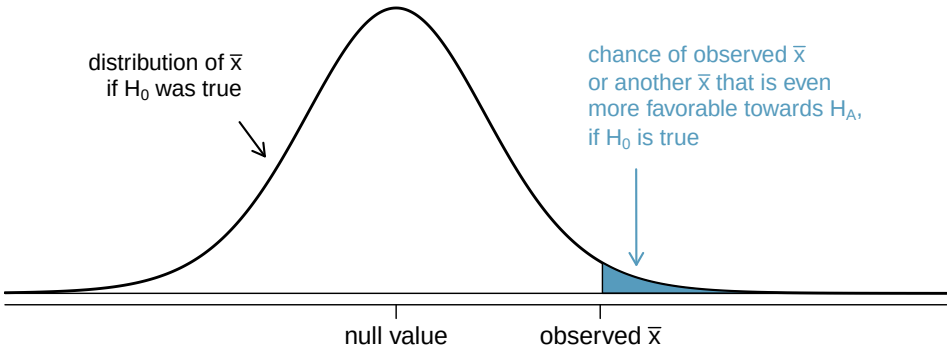


Figure 8.11: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favorable to H_a under this distribution.

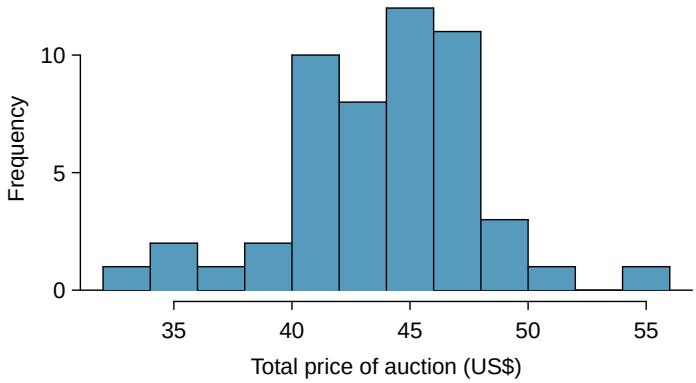


Figure 8.12: A histogram of the total auction prices for 52 Ebay auctions.

⊙ **Exercise 8.12** If the null hypothesis is true, how often should the p-value be less than 0.05?[13]

⊙ **Exercise 8.13** Suppose we had used a significance level of 0.01 in the sleep study. Would the evidence have been strong enough to reject the null hypothesis? (The p-value was 0.007.) What if the significance level was $\alpha = 0.001$? [14]

⊙ **Exercise 8.14** Ebay might be interested in showing that buyers on its site tend to pay less than they would for the corresponding new item on Amazon. We'll research this topic for one particular product: a video game called *Mario Kart* for the Nintendo Wii. During early October 2009, Amazon sold this game for $46.99. Set up an appropriate (one-sided!) hypothesis test to check the claim that Ebay buyers pay less during auctions at this same time.[15]

⊙ **Exercise 8.15** During early October, 2009, 52 Ebay auctions were recorded for *Mario Kart*.[16] The total prices for the auctions are presented using a histogram in Figure 8.12, and we may like to apply the normal model to the sample mean. Check the three conditions required for applying the normal model: (1) independence, (2) at least 30 observations, and (3) the data are not strongly skewed.[17]

---

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

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

[15] The skeptic would say the average is the same on Ebay, and we are interested in showing the average price is lower.

H_0: The average auction price on Ebay is equal to (or more than) the price on Amazon. We write only the equality in the statistical notation: $\mu_{ebay} = 46.99$.

H_a: The average price on Ebay is less than the price on Amazon, $\mu_{ebay} < 46.99$.

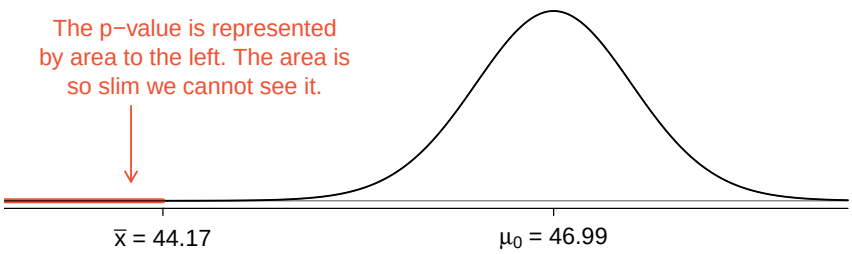
[16] These data were collected by OpenIntro staff.

[17] (1) The independence condition is unclear. *We will make the assumption that the observations are independent, which we should report with any final results.* (2) The sample size is sufficiently large: $n = 52 \geq 30$. (3) The data distribution is not strongly skewed; it is approximately symmetric.

Example 8.16 The average sale price of the 52 Ebay auctions for *Wii Mario Kart* was \$44.17 with a standard deviation of \$4.15. Does this provide sufficient evidence to reject the null hypothesis in Exercise (8.14)? Use a significance level of $\alpha = 0.01$.

The hypotheses were set up and the conditions were checked in Exercises (8.14) and (8.15). The next step is to find the standard error of the sample mean and produce a sketch to help find the p-value.

$$SE_{\bar{x}} = s/\sqrt{n} = 4.15/\sqrt{52} = 0.5755$$



Because the alternative hypothesis says we are looking for a smaller mean, we shade the lower tail. We find this shaded area by using the Z score and normal probability table: $Z = \frac{44.17-46.99}{0.5755} = -4.90$, which has area less than 0.0002. The area is so small we cannot really see it on the picture. This lower tail area corresponds to the p-value.

Because the p-value is so small – specifically, smaller than $\alpha = 0.01$ – this provides sufficiently strong evidence to reject the null hypothesis in favor of the alternative. The data provide statistically significant evidence that the average price on Ebay is lower than Amazon's asking price.

8.3.1 Two-sided hypothesis testing with p-values

We now consider how to compute a p-value for a two-sided test. In one-sided tests, we shade the single tail in the direction of the alternative hypothesis. For example, when the alternative had the form $\mu > 7$, then the p-value was represented by the upper tail (Figure 8.11). When the alternative was $\mu < 46.99$, the p-value was the lower tail (Exercise (8.14)). In a two-sided test, *we shade two tails* since evidence in either direction is favorable to H_a.

⊙ **Exercise 8.17** Earlier we talked about a research group investigating whether the students at their school slept longer than 7 hours each night. Let's consider a second group of researchers who want to evaluate whether the students at their college differ from the norm of 7 hours. Write the null and alternative hypotheses for this investigation.[18]

● **Example 8.18** The second college randomly samples 72 students and finds a mean of $\bar{x} = 6.83$ hours and a standard deviation of $s = 1.8$ hours. Does this provide strong evidence against H_0 in Exercise (8.17)? Use a significance level of $\alpha = 0.05$.

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

Next we can compute the standard error ($SE_{\bar{x}} = \frac{s}{\sqrt{n}} = 0.21$) of the estimate and create a picture to represent the p-value, shown in Figure 8.13. Both tails are shaded. An estimate of 7.17 or more provides at least as strong of evidence against the null hypothesis and in favor of the alternative as the observed estimate, $\bar{x} = 6.83$.

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

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

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

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

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

● **Example 8.19** It is never okay to change two-sided tests to one-sided tests after observing the data. In this example we explore the consequences of ignoring this advice. Using $\alpha = 0.05$, we show that freely switching from two-sided tests to one-sided tests will cause us to make twice as many Type 1 Errors as intended.

Suppose the sample mean was larger than the null value, μ_0 (e.g. μ_0 would represent 7 if H_0: $\mu = 7$). Then if we can flip to a one-sided test, we would use H_a: $\mu > \mu_0$. Now

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

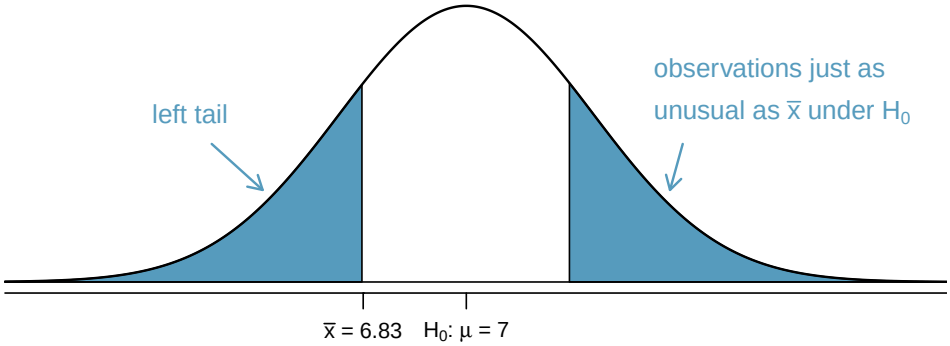


Figure 8.13: H_a is two-sided, so *both* tails must be counted for the p-value.

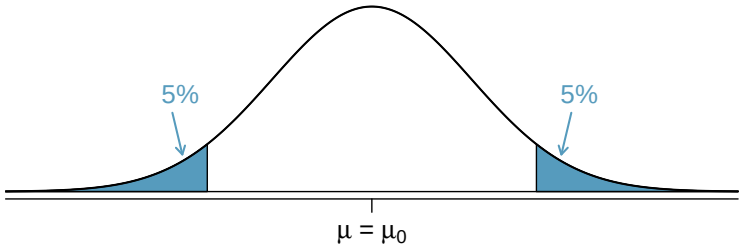


Figure 8.14: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example (8.19) when $\alpha = 0.05$.

if we obtain any observation with a Z score greater than 1.65, we would reject H_0. If the null hypothesis is true, we incorrectly reject the null hypothesis about 5% of the time when the sample mean is above the null value, as shown in Figure 8.14.

Suppose the sample mean was smaller than the null value. Then if we change to a one-sided test, we would use H_a: $\mu < \mu_0$. If $\bar{x}$ had a Z score smaller than -1.65, we would reject H_0. If the null hypothesis is true, then we would observe such a case about 5% of the time.

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

Caution: One-sided hypotheses are allowed only *before* seeing data
After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Hypotheses must be set up *before* observing the data. If they are not, the test must be two-sided.

8.4 Choosing a significance level (special topic)

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

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 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 should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false. We will discuss this particular case in greater detail in Section 6.4.

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

Example 8.20 A car manufacturer is considering a higher quality but more expensive supplier for window parts in its vehicles. They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier. Is there good reason to modify the significance level in such a hypothesis test?

The null hypothesis is that the more expensive parts last no more than 12% longer while the alternative is that they do last more than 12% longer. This decision is just one of the many regular factors that have a marginal impact on the 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 8.21 The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not windows. 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?

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.

Exercise 8.22 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.[19]

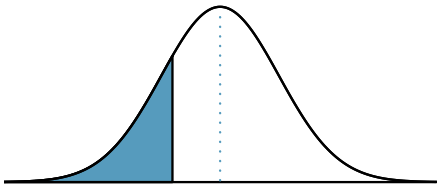
[19] Here the null hypothesis is that the part is not broken, and the alternative is that it is broken. If we

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 $\alpha = 0.01$.

Appendix A

Distribution tables

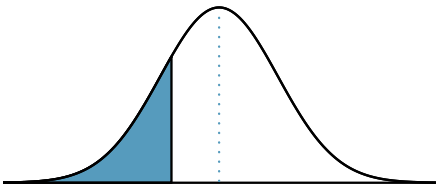
Normal Probability Table



The area to the left of Z represents the percentile of the observation. The normal probability table always lists percentiles

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.

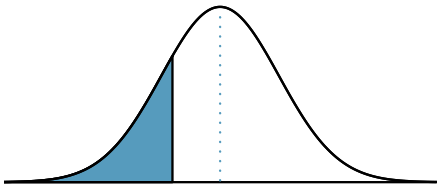


The area to the left of Z represents the percentile of the observation. The normal probability table always lists percentiles

	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.

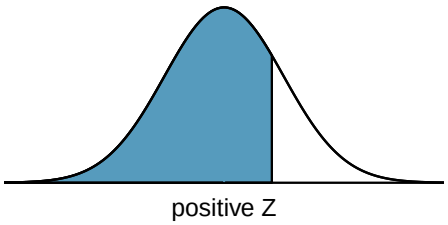
A.1 Standard Normal Probability Table



The area to the left of Z represents the percentile of the observation.

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.



The area to the left of Z represents the percentile of the observation.

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
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.

A.2 t Distribution Table

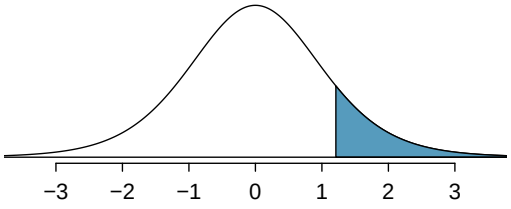


Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

		Upper-tail probability p											
		0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
df	1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
	2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
	3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
	4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
	5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
	6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
	7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
	8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
	9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
	10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
	11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
	12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
	13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
	14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
	15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
	16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
	17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
	18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
	19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
	20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
	21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
	22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
	23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
	24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
	25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
	26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
	27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
	28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
	29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
	30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
	40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
	50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
	60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
	80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
	100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
	1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
	z^*	0.674	0.841	1.036	1.282	1.645	1.96	2.054	2.326	2.576	2.807	3.091	3.291
		50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
		Confidence Level C											

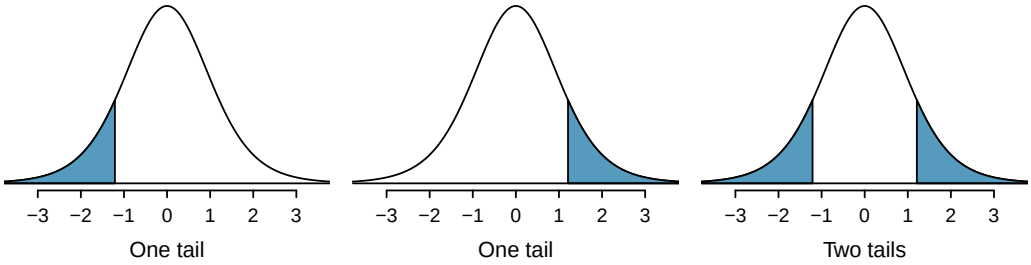


Figure A.1: Three t distributions.

one tail		0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two tails		0.40	0.30	0.20	0.10	0.050	0.020	0.01	0.002	0.0010
df	31	0.853	1.054	1.309	1.696	2.040	2.453	2.744	3.375	3.633
	32	0.853	1.054	1.309	1.694	2.037	2.449	2.738	3.365	3.622
	33	0.853	1.053	1.308	1.692	2.035	2.445	2.733	3.356	3.611
	34	0.852	1.052	1.307	1.691	2.032	2.441	2.728	3.348	3.601
	35	0.852	1.052	1.306	1.690	2.030	2.438	2.724	3.340	3.591
	36	0.852	1.052	1.306	1.688	2.028	2.434	2.719	3.333	3.582
	37	0.851	1.051	1.305	1.687	2.026	2.431	2.715	3.326	3.574
	38	0.851	1.051	1.304	1.686	2.024	2.429	2.712	3.319	3.566
	39	0.851	1.050	1.304	1.685	2.023	2.426	2.708	3.313	3.558
	40	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
	45	0.850	1.049	1.301	1.679	2.014	2.412	2.690	3.281	3.520
	50	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261	3.496
	55	0.848	1.046	1.297	1.673	2.004	2.396	2.668	3.245	3.476
	60	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
	65	0.847	1.045	1.295	1.669	1.997	2.385	2.654	3.220	3.447
	70	0.847	1.044	1.294	1.667	1.994	2.381	2.648	3.211	3.435
	75	0.846	1.044	1.293	1.665	1.992	2.377	2.643	3.202	3.425
	80	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
	85	0.846	1.043	1.292	1.663	1.988	2.371	2.635	3.189	3.409
	90	0.846	1.042	1.291	1.662	1.987	2.368	2.632	3.183	3.402
	100	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
	200	0.843	1.039	1.286	1.653	1.972	2.345	2.601	3.131	3.340
	250	0.843	1.039	1.285	1.651	1.969	2.341	2.596	3.123	3.330
	300	0.843	1.038	1.284	1.650	1.968	2.339	2.592	3.118	3.323
	350	0.843	1.038	1.284	1.649	1.967	2.337	2.590	3.114	3.319
	400	0.843	1.038	1.284	1.649	1.966	2.336	2.588	3.111	3.315
	500	0.842	1.038	1.283	1.648	1.965	2.334	2.586	3.107	3.310
	600	0.842	1.037	1.283	1.647	1.964	2.333	2.584	3.104	3.307
	750	0.842	1.037	1.283	1.647	1.963	2.331	2.582	3.101	3.304
	1000	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
	2000	0.842	1.037	1.282	1.646	1.961	2.328	2.578	3.094	3.295
	3000	0.842	1.037	1.282	1.645	1.961	2.328	2.577	3.093	3.294
	4000	0.842	1.037	1.282	1.645	1.961	2.327	2.577	3.092	3.293
	5000	0.842	1.037	1.282	1.645	1.960	2.327	2.577	3.092	3.292
	∞	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

A.3 t Distribution Table

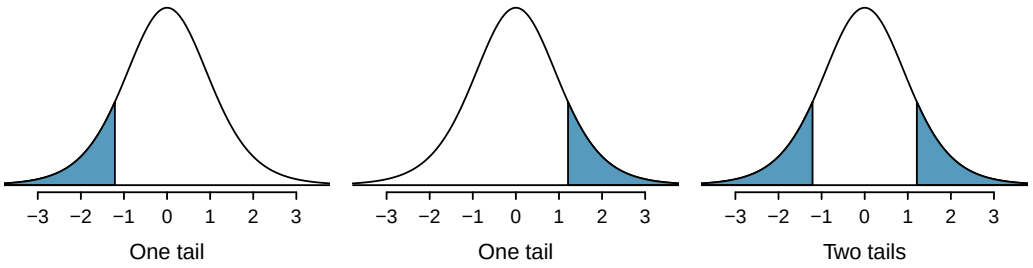


Figure A.2: Three t distributions.

one tail	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two tails	0.40	0.30	0.20	0.10	0.050	0.020	0.01	0.002	0.0010
df 1	1.376	1.963	3.078	6.314	12.710	31.820	63.660	318.300	636.600
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.330	31.600
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.210	12.920
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.906	1.134	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646

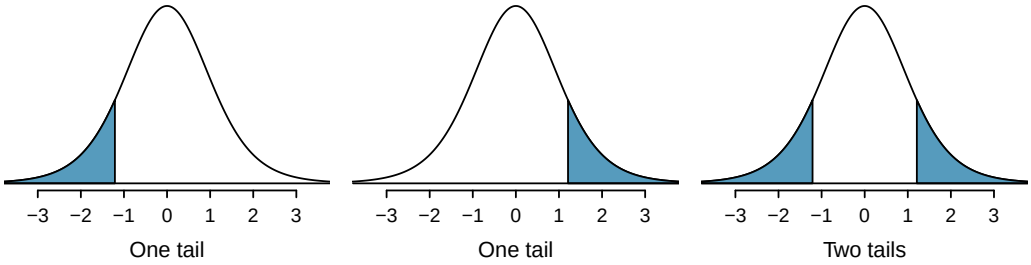


Figure A.3: Three t distributions.

one tail		0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two tails		0.40	0.30	0.20	0.10	0.050	0.020	0.01	0.002	0.0010
df	31	0.853	1.054	1.309	1.696	2.040	2.453	2.744	3.375	3.633
	32	0.853	1.054	1.309	1.694	2.037	2.449	2.738	3.365	3.622
	33	0.853	1.053	1.308	1.692	2.035	2.445	2.733	3.356	3.611
	34	0.852	1.052	1.307	1.691	2.032	2.441	2.728	3.348	3.601
	35	0.852	1.052	1.306	1.690	2.030	2.438	2.724	3.340	3.591
	36	0.852	1.052	1.306	1.688	2.028	2.434	2.719	3.333	3.582
	37	0.851	1.051	1.305	1.687	2.026	2.431	2.715	3.326	3.574
	38	0.851	1.051	1.304	1.686	2.024	2.429	2.712	3.319	3.566
	39	0.851	1.050	1.304	1.685	2.023	2.426	2.708	3.313	3.558
	40	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
	45	0.850	1.049	1.301	1.679	2.014	2.412	2.690	3.281	3.520
	50	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261	3.496
	55	0.848	1.046	1.297	1.673	2.004	2.396	2.668	3.245	3.476
	60	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
	65	0.847	1.045	1.295	1.669	1.997	2.385	2.654	3.220	3.447
	70	0.847	1.044	1.294	1.667	1.994	2.381	2.648	3.211	3.435
	75	0.846	1.044	1.293	1.665	1.992	2.377	2.643	3.202	3.425
	80	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
	85	0.846	1.043	1.292	1.663	1.988	2.371	2.635	3.189	3.409
	90	0.846	1.042	1.291	1.662	1.987	2.368	2.632	3.183	3.402
	100	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
	200	0.843	1.039	1.286	1.653	1.972	2.345	2.601	3.131	3.340
	250	0.843	1.039	1.285	1.651	1.969	2.341	2.596	3.123	3.330
	300	0.843	1.038	1.284	1.650	1.968	2.339	2.592	3.118	3.323
	350	0.843	1.038	1.284	1.649	1.967	2.337	2.590	3.114	3.319
	400	0.843	1.038	1.284	1.649	1.966	2.336	2.588	3.111	3.315
	500	0.842	1.038	1.283	1.648	1.965	2.334	2.586	3.107	3.310
	600	0.842	1.037	1.283	1.647	1.964	2.333	2.584	3.104	3.307
	750	0.842	1.037	1.283	1.647	1.963	2.331	2.582	3.101	3.304
	1000	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
	2000	0.842	1.037	1.282	1.646	1.961	2.328	2.578	3.094	3.295
	3000	0.842	1.037	1.282	1.645	1.961	2.328	2.577	3.093	3.294
	4000	0.842	1.037	1.282	1.645	1.961	2.327	2.577	3.092	3.293
	5000	0.842	1.037	1.282	1.645	1.960	2.327	2.577	3.092	3.292
	∞	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291